

# Bootstrapping and Sample Splitting For High-Dimensional, Assumption-Free Inference

Alessandro Rinaldo, Larry Wasserman, Max G'Sell, Jing Lei and Ryan Tibshirani

Carnegie Mellon University

November 16, 2016

*Several new methods have been proposed for performing valid inference after model selection. An older method is sampling splitting: use part of the data for model selection and part for inference. In this paper we revisit sample splitting combined with the bootstrap (or the Normal approximation). We show that this leads to a simple, assumption-free approach to inference and we establish results on the accuracy of the method. In fact, we find new bounds on the accuracy of the bootstrap and the Normal approximation for general nonlinear parameters with increasing dimension which we then use to assess the accuracy of regression inference. We show that an alternative, called the image bootstrap, has higher coverage accuracy at the cost of more computation. We define new parameters that measure variable importance and that can be inferred with greater accuracy than the usual regression coefficients. There is a inference-prediction tradeoff: splitting increases the accuracy and robustness of inference but can decrease the accuracy of the predictions.*

“Investigators who use [regression] are not paying adequate attention to the connection - if any - between the models and the phenomena they are studying. ... By the time the models are deployed, the scientific position is nearly hopeless. Reliance on models in such cases is Panglossian ...”

—David Freedman

## 1 Introduction

We consider the problem of carrying out assumption-free statistical inference after model selection for high-dimensional linear regression. This is now a large topic and a variety of approaches have been considered under different settings; see, for example, [Buja et al. \(2015\)](#); [Leeb and Pötscher \(2008\)](#); [Meinshausen et al. \(2012\)](#); [Lei et al. \(2016\)](#); [Lockhart et al. \(2014\)](#); [Lee et al. \(2016\)](#); [Taylor et al. \(2014\)](#); [Berk et al. \(2013\)](#); [Javanmard and Montanari \(2014\)](#); [Mentch and Hooker \(2016\)](#).

In this paper, we will use linear models but we do not assume that the true regression function is linear. We show the following:

1. Inference based on sample splitting followed by the bootstrap (or Normal approximation) gives assumption-free, robust confidence intervals under very weak assumptions. No other known method gives the same inferential guarantees.
2. The usual regression parameters are not the best choice of parameter to estimate in the weak assumption case. We propose new parameters, called LOCO (Leave-Out-COvariates) parameters, that are interpretable, general and can be estimated accurately.
3. There is a tradeoff between prediction accuracy and inferential accuracy.

4. We provide new bounds on the accuracy of the Normal approximation and the bootstrap to the distribution of the projection parameter (the best linear predictor) when the dimension increases and the model is wrong. We need these bounds since we will use Normal approximations or the bootstrap after choosing the model. In fact, we provide new general bounds on Normal approximations for nonlinear parameters with increasing dimension. This gives new insights on the accuracy of inference in high-dimensional situations. In particular, the accuracy of the Normal approximation for the standard regression parameters is very poor while the approximation is very good for LOCO parameters.
5. The accuracy of the bootstrap can be improved by using an alternative version that we call the image bootstrap. However, this version is computationally expensive.
6. We show that the law of the projection parameter cannot be consistently estimated without sample splitting.

**The Model.** We observe iid pairs  $Z_1, \dots, Z_n \sim P$  where  $Z_i = (X_i, Y_i)$ ,  $X_i \in \mathbb{R}^d$  and  $Y_i \in \mathbb{R}$ . The regression function  $\mu(x) = \mathbb{E}[Y|X = x]$  is not assumed to be linear. We use the data to choose a subset of covariates  $S \subset \{1, \dots, d\}$ ; thus  $S = w_n(Z_1, \dots, Z_n)$  for some selection function  $w_n$  such that  $k_n \equiv |S| < n$ . We choose a parameter  $\theta_S$  that depends on  $S$ . For example,  $\theta_S$  could be the coefficients of the best linear predictor based on the selected covariates  $X_S = (X(j) : j \in S)$ . The parameter  $\theta_S$  is random because it depends on the selected variables  $S$ . We want a confidence set  $C_S$  that satisfies

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} P^n(\theta_S \in C_S) \geq 1 - \alpha \quad (1)$$

where  $\mathcal{P}_n$  is a large, nonparametric class of distributions. In other words,  $C_S$  is an *honest confidence set* (Li, 1989). In fact, our confidence sets will be *robust*, honest confidence sets, which means that

$$\liminf_{n \rightarrow \infty} \inf_{w_n \in \mathcal{W}_n} \inf_{P \in \mathcal{P}_n} P^n(\theta_S \in C_S) \geq 1 - \alpha \quad (2)$$

where  $\mathcal{W}_n$  is the set of all selection functions. The reason we want validity over all selection rules is because, in realistic data analysis, the selection rule can be very complex. The choice of model can involve: plotting, outlier removal, transformations, trying various models etc. There is room for unconscious biases to enter unless we have validity over all  $w_n$ .

To the best of our knowledge there are only two methods that achieve honest coverage: sample splitting and uniform inference. Uniform inference is based on estimating the distribution of the parameter estimates over all possible model selections. In general, this is infeasible. But we compare sample splitting and uniform inference in a restricted model in Section 7.

The confidence sets we construct will be hyper-rectangles. The reason we use rectangles is two-fold. First, once we have a rectangular confidence set for a vector parameter, we immediately have simultaneous confidence intervals for the components of the vector. Second, the high dimensional central limit theorem is more accurate for rectangles than for other shapes (Chernozhukov et al., 2013, 2014).

We allow  $d = d_n$  to increase with  $n$ . There is no restriction on  $d$ . We assume that we have a model

selection algorithm which chooses a subset  $S$  of variables with  $|S| \leq k \equiv k_n < n$ . We consider the following three models for  $Z = (X, Y)$ :

$$\mathcal{P}_n = \left\{ \text{all distributions on } \mathbb{R}^{d+1} \right\} \quad (3)$$

$$\mathcal{P}'_n = \left\{ P : \text{support}(P) \subset [-A, A]^{d+1} \right\} \quad (4)$$

$$\mathcal{P}^\dagger_n = \left\{ P : \text{support}(P) \subset [-A, A]^{d+1}, \min_j \text{Var}_P(W_i(j)) \geq a, \min_{|S| \leq k} \lambda_{\min}(\Sigma_S) \geq u, \right. \\ \left. \max_{|S| \leq k} \lambda_{\max}(\Sigma_S) \leq U, \min_{|S| \leq k} \lambda_{\min}(V(P)) \geq v, P \text{ has a density} \right\} \quad (5)$$

where  $W = (\text{vech}(XX^T), XY)$ ,  $\Sigma_S = \mathbb{E}[X_S X_S^T]$ ,  $Z_S = (Y, X_S)$ ,  $V(P) = \text{Var}(W)$ ,  $a, A, u, U, v$  are (unknown) finite positive constants and  $\lambda_{\min}$  denotes the minimum eigenvalue. We will define certain parameters  $\beta_S$ ,  $\phi_S$ ,  $\gamma_S$  and  $\rho_S$ . For inferences on  $\phi_S$ , we use the first model, for inferences on  $\gamma_S$  and  $\rho_S$  we use the second, and for inferences on  $\beta_S$  we use the third model.

In all cases, the mean function  $\mu(x) = \mathbb{E}[Y|X = x]$  is unrestricted. The compact support assumption can be replaced with moment bounds. We make the assumptions slightly stronger than necessary to keep the statement of the results and their proofs simpler. The above models place no restriction on the model selection procedure or on  $d$ .

**Sample Splitting.** The oldest method for inference after model selection is sample splitting: half the data  $\mathcal{D}_1$  are used for model fitting and the other half  $\mathcal{D}_2$  are used for inference.<sup>1</sup> Thus  $S = w_{n/2}(\mathcal{D}_1)$ . The earliest references for sample splitting that we know of are [Barnard \(1974\)](#), [Cox \(1975\)](#), [Faraway \(1995\)](#) [Hartigan \(1969\)](#), page 13 of [Miller \(1990\)](#) [Moran \(1973\)](#), page 37 of [Mosteller and Tukey \(1977\)](#) and [Picard and Berk \(1990\)](#). To quote Barnard: “... the simple idea of splitting a sample in two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics ...”

## 1.1 Goals

The purpose of this paper is to formally analyze inference based on sample splitting and Normal (or bootstrap) approximations. We want the estimator  $\hat{\theta}_S$  and confidence set  $C_S$  to have the following properties:

---

<sup>1</sup> For simplicity, we assume that the data are split into two parts of equal size. The problem of determining the optimal size of the split is not considered in this paper.

$$\text{Coverage validity (honesty)} : \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}} P^n(\theta_S \in C_S) \geq 1 - \alpha \quad (6)$$

$$\text{Concentration} : \sup_{P \in \mathcal{P}} P^n(\|\hat{\theta}_S - \theta_S\|_\infty > r_n) \rightarrow 0 \quad (7)$$

$$\text{Accuracy of } C_n : \sup_{P \in \mathcal{P}} P^n(\nu(C_S) > \epsilon_n) \rightarrow 0 \quad (8)$$

where  $\mathcal{P} \in \{\mathcal{P}_n, \mathcal{P}'_n, \mathcal{P}^\dagger_n\}$ ,  $r_n, \epsilon_n = o(1)$  and  $\nu$  is Lebesgue measure. In the probability statements above, both the parameter and the confidence set are random. Generally,  $\theta_S$  is a vector of size  $|S|$ . As mentioned earlier, we will actually get a stronger guarantee than honesty; inferences will be honest and robust:

$$\liminf_{n \rightarrow \infty} \inf_{w \in \mathcal{W}_n} \inf_{P \in \mathcal{P}_n} P^n(\theta_S \in C_S) \geq 1 - \alpha. \quad (9)$$

## 1.2 Parameters For Regression

For any fixed subset of variables  $S \subset \{1, \dots, d\}$  let  $X_S = (X(j) : j \in S)$  be the corresponding set of covariates. The *projection parameter*  $\beta_S \in \mathbb{R}^{|S|}$  is defined to be the coefficient vector of the best linear predictor using  $X_S$ . In other words,  $\beta_S$  minimizes  $\mathbb{E}_{X,Y}(Y - \beta^T X_S)^2$  where  $(X, Y)$  denotes a new pair. This type of parameter is used in [Lee et al. \(2016\)](#); [Taylor et al. \(2014\)](#); [Berk et al. \(2013\)](#); [Wasserman \(2014\)](#). Note that  $\beta_S$  is well-defined even though we do not assume  $\mu(x)$  is linear. We have that

$$\beta_S = \Sigma_S^{-1} \alpha_S \quad (10)$$

where  $\alpha_S = (\alpha_S(j) : j \in S)$ ,  $\alpha_S(j) = \mathbb{E}[Y X_S(j)]$  and  $\Sigma_S = \mathbb{E}[X_S X_S^T]$ . We write  $\beta_S = g(\psi)$  where  $\psi = (\sigma, \alpha)$  and  $\sigma = \text{vec}(\Sigma)$ . When  $S$  is selected using the data,  $\beta_S$  is a randomly selected parameter.

Often, statisticians are interested in  $\beta_S$  as a measure of the importance of the selected covariates. But there are other ways to measure variable importance as well. We now define several *Leave Out COvariate Inference (LOCO) parameters*. These were also defined in [Lei et al. \(2016\)](#) and are similar to the variable importance measures used in random forests. The first is  $\gamma_S = (\gamma_S(j) : j \in S)$  where

$$\gamma_S(j) = \mathbb{E}_{X,Y} \left[ |Y - \hat{\beta}_{S(j),j}^T X_S| - |Y - \hat{\beta}_S^T X_S| \right],$$

$Z = (X, Y) \sim P$  is a new pair, the expectation is only over  $(X, Y)$  only,  $\hat{\beta}_S$  is any estimator of  $\beta_S$  and  $\hat{\beta}_{S(j),j}$  is obtained by re-running the model selection and estimation procedure after removing the  $j^{\text{th}}$  covariate. We write  $S(j)$  since the selected model can be different when covariate  $j$  is held out from the data. Thus,  $\gamma_S(j)$  has a very simple interpretation: it is the increase in prediction error by not having access to  $X(j)$ . Of course, we can define a version that leaves out several variables at once if we like. The parameter  $\gamma_S$  has several advantages over the projection parameter: it is more interpretable since it refers directly to prediction error and we shall see that the accuracy of

the Normal approximation and the bootstrap is much higher. Indeed, we believe that the current focus on  $\beta_S$  is mainly due to the fact that statisticians are used to thinking in terms of cases where the linear model is assumed to be correct. Another LOCO parameter that we consider is  $\phi_S = (\phi_S(j) : j \in S)$  where

$$\phi_S(j) = \text{median} \left[ |Y - \hat{\beta}_{S(j),j}^T X_S| - |Y - \hat{\beta}_S^T X_S| \right], \quad (11)$$

where the median is over the distribution of  $(X, Y)$ . It is also of interest to have a parameter that measures how well the selected model will predict future observations. To this end, we define the future predictive error

$$\rho_S = \mathbb{E}_{X,Y} \left[ |Y - \hat{\beta}_S^T X_S| \right]. \quad (12)$$

In general, we let  $\theta_S$  denote the parameter of interest which, depending on the context, can be  $\beta_S$ ,  $\gamma_S$ ,  $\phi_S$  or  $\rho_S$ .

**Remark:** The LOCO parameters do not require linear models. For example we can define

$$\gamma_S(j) = \mathbb{E}_{X,Y} \left[ |Y - \hat{\mu}_j(X_{S(j)})| - |Y - \hat{\mu}(X_S)| \right]$$

where  $\hat{\mu}$  is any regression estimator and  $\hat{\mu}_j$  is the estimator obtained after refitting without covariate  $j$ . However, we will focus on linear models in this paper.

**Remark.** There is another version of the projection parameter defined as follows. For the moment, suppose that  $d < n$  and that there is no model selection. Let  $\beta_n = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mu_n$  where  $\mathbb{X}$  is the  $n \times d$  design matrix,  $\mu_n = (\mu_n(1), \dots, \mu_n(n))^T$  and  $\mu_n(i) = \mathbb{E}[Y_i | X_1, \dots, X_n]$ . This is just the mean of the least squares estimator conditional on  $X_1, \dots, X_n$ . We call this the *conditional projection parameter*. The meaning of this parameter when the linear model is false, is not clear. It is a data dependent parameter, even in the absence of model selection. [Buja et al. \(2015\)](#) have devoted a whole paper to this issue. Quoting from their paper:

When fitted models are approximations, conditioning on the regressor is no longer permitted ... Two effects occur: (1) parameters become dependent on the regressor distribution; (2) the sampling variability of the parameter estimates no longer derives from the conditional distribution of the response alone. Additional sampling variability arises when the nonlinearity conspires with the randomness of the regressors to generate a  $1/\sqrt{n}$  contribution to the standard errors.

Moreover, it is not possible to estimate the distribution of the conditional projection parameter estimate in the distribution free framework. To see that, note that the least squares estimator can be written as  $\hat{\beta}(j) = \sum_{i=1}^n w_i Y_i$  for weights  $w_i$  that depend on the design matrix. Then  $\sqrt{n}(\hat{\beta}(j) - \beta(j)) = \sum_{i=1}^n w_i \epsilon_i$  where  $\epsilon_i = Y_i - \mu_n(i)$ . Thus,  $\sqrt{n}(\hat{\beta}(j) - \beta(j)) \approx N(0, \tau^2)$  where  $\tau^2 = \sum_i w_i^2 \sigma_i^2$  where  $\sigma_i^2 = \text{Var}(\epsilon_i | X_1, \dots, X_n)$ . The problem is that there is no consistent estimator

of  $\tau^2$  under the nonparametric models we are considering. Even if we assume that  $\sigma_i^2$  is constant (an assumption we avoid in this paper), we still have that  $\tau^2 = \sigma^2 \sum_i w_i^2$  which cannot be consistently estimated without assuming that the linear model is correct. Again, we refer the reader to [Buja et al. \(2015\)](#) for more discussion.

In contrast, the projection parameter defined in (10) is  $\beta = \Sigma^{-1}\alpha$  which is a fixed functional  $T(P)$  of the distribution  $P$  and the distribution of the estimator is estimable. For these reasons, we focus in this paper on the projection parameter rather than the conditional projection parameter.

### 1.3 Coverage Accuracy

We construct inferences using  $\mathcal{D}_2$ . The inferences are based on either a Normal approximation or the bootstrap. This raises the question: how accurate are the confidence sets given that we do not assume that the linear model is correct? We will show (Theorem 9) that, for the projection parameter, the coverage of the Normal or bootstrap confidence set  $R_n$  is given by

$$\inf_{P \in \mathcal{P}_n^\dagger} P^n(\beta \in R_n) = 1 - \alpha - O(\Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3}) \quad (13)$$

where

$$\begin{aligned} \Delta_{n,1} &= \max \left\{ \frac{1}{n^{1/6}}, \frac{1}{\sqrt{v}} \left( \frac{k^6 (\log kn)^7}{n} \right)^{1/6} \right\} \\ \Delta_{n,2} &= \frac{U}{\sqrt{v}} \sqrt{\frac{k^6 \log^2 n \log k}{n u^6}} \\ \Delta_{n,3} &= \left( \frac{U}{\sqrt{v}} \right)^{1/3} \left( \frac{k^5 \log k + \log n}{u^{10} n} \log^4 k \right)^{1/6}. \end{aligned}$$

The constants  $U$  and  $v$  are defined in (18). The term  $\Delta_{n,1}$  comes from the high dimensional Berry-Esseen theorem. The term  $\Delta_{n,2}$  comes from the fact that the projection parameter is non-linear. The term  $\Delta_{n,3}$  is due to the fact that the covariance of the estimator is unknown. For the Normal approximation, we have to estimate this covariance. The bootstrap implicitly estimates this covariance. The dominant term of the inaccuracy is of order  $k/(n^{1/6}u^3)$ . But for the LOCO parameter  $\gamma_S$ , we have the more accurate coverage

$$\inf_{P \in \mathcal{P}_n'} P^n(\gamma_S \in R_n) = 1 - \alpha - O \left( \frac{(\log kn)^7}{n} \right)^{1/6} \quad (14)$$

and for  $\phi_S$ ,

$$\inf_{P \in \mathcal{P}_n} P^n(\phi_S \in R_n) \geq 1 - \alpha. \quad (15)$$

A few remarks are in order:

1. These are upper bounds. At this time, we do not know if they are sharp although we believe they are.

2. The dependence on dimension and on the minimal eigenvalue is quite poor. We believe this is unavoidable and is one reason we recommend using LOCO parameters instead of projection parameters. Our findings are consistent with [El Karoui and Purdom \(2015\)](#) who found, even when the linear model is correct, that the bootstrap does poorly when  $k_n$  increases.
3. As a function of sample size, there is a term of order  $n^{-1/6}$  in  $\Delta_n$ . The exponent  $1/6$  comes from the Berry-Esseen bound in Section 3. [Chernozhukov et al. \(2014\)](#) conjecture that this rate is optimal for high-dimensional central limit theorems. Their conjecture is based on the lower bound result in [Bentkus \(1985\)](#). If their conjecture is true, then this is best rate that can be hoped for in general.
4. The rate is much slower than the rate obtained in the central limit theorem given in [Portnoy \(1987\)](#) for robust regression estimators. A reason for such discrepancy is that [Portnoy \(1987\)](#) assumes, among the other things, that the linear model is correct. In this case, the least squares estimators is conditionally unbiased. Without the assumption of model correctness there is a substantial bias.
5. If we assume that the covariates are independent then the situation gets dramatically better. For example, the term  $\Delta_{n,2}$  is then  $O(1/\sqrt{n})$ . But the goal of this paper is to avoid adding such assumptions.

The dependence of the accuracy of inferences for the projection parameter on the minimum eigenvalue  $u$  is common and seems unavoidable. Realistically, if  $d$  is huge, then we have probably collected many redundant covariates and we expect  $u \approx 0$ . The dependence on the dimension  $k$  is also very strong. We regard this as another reason to prefer the LOCO parameters. The coverage accuracy can also be improved by changing the bootstrap procedure; see Section 5.

## 1.4 Related Work

The problem of inference after model selection has received much attention lately. [Lockhart et al. \(2014\)](#); [Lee et al. \(2016\)](#); [Taylor et al. \(2014\)](#) have a series of interesting papers developing a conditional approach. Basically, the distribution of  $\sqrt{n}(\hat{\beta}(j) - \beta(j))$  conditional on the selected model, has a truncated Gaussian distribution if the errors are Normal and the covariates are fixed. The cdf of the truncated Gaussian is used as a pivot to get tests and confidence intervals. This approach requires Normality, and a fixed, known variance. [Berk et al. \(2013\)](#) use a simultaneous inference approach, taking a max over all possible model selections. The method also assumes Normality and a fixed, known variance. It is also computationally very expensive. [Javanmard and Montanari \(2014\)](#) construct confidence intervals for high dimensional regression. These can be used for the selected model if a Bonferroni correction is applied. See also [Nickl et al. \(2013\)](#). However, these methods assume that the linear model is correct as well as a number of other assumptions. We discuss these approaches in more detail in Section 8.

## 1.5 Outline

In Section 2 we introduce the basic sample splitting strategies. In Section 3 we establish a Berry-Esseen bound for regression with possibly increasing dimension and no assumption of linearity on

the regression function. In Section 4 we establish the accuracy of the estimators and the confidence sets. In Section 5 we briefly discuss a few strategies for improving the accuracy of the inference. In Section 6 we discuss the variability due to splitting. In Section 7 we compare sample splitting to non-splitting strategies. Section 8 contains some comments on other methods. In Section 9 we report some numerical examples. Finally, Section 10 contains concluding remarks.

## 1.6 Notation

Let  $Z = (X, Y) \sim P$  where  $Y \in \mathbb{R}$  and  $X \in \mathbb{R}^d$ . We write  $X = (X(1), \dots, X(d))$  to denote the components of the vector  $X$ . Define  $\Sigma = \mathbb{E}[XX^T]$  and  $\alpha = (\alpha(1), \dots, \alpha(d))$  where  $\alpha(j) = \mathbb{E}[YX(j)]$ . Let  $\sigma = \text{vec}(\Sigma)$  and  $\psi \equiv \psi(P) = (\sigma, \alpha)$ . The regression function is  $\mu(x) = \mathbb{E}[Y|X = x]$ . We use  $\nu$  to denote Lebesgue measure. We write  $a_n \preceq b_n$  to mean that there exists a constant  $C > 0$  such that  $a_n \leq Cb_n$  for all large  $n$ .  $S \subset \{1, \dots, d\}$  denotes a subset of the covariates and  $X_S = (X(j) : j \in S)$  denotes the corresponding elements of  $X$ . Similarly,  $\Sigma_S = \mathbb{E}[X_S X_S^T]$  and  $\alpha_S = \mathbb{E}[YX_S]$ . Let  $Z_S = (X_S, Y)$ . We write  $\Omega = \Sigma^{-1}$  and  $\omega = \text{vec}(\Omega)$  where  $\text{vec}$  is the operator that stacks a matrix into one large vector.  $A \otimes B$  denotes the kronecker product of matrices. The commutation matrix  $T_{m,n}$  is the  $mn \times mn$  matrix defined by  $T_{m,n} \text{vec}(A) = \text{vec}(A^T)$  for any  $m \times n$  matrix  $A$ .  $\text{vech}(A)$  denotes the column vector of dimension  $k(k+1)/2$  obtained by vectorizing only the lower triangular part of  $k \times k$  matrix  $A$ .

## 2 Sample Splitting

We split the data into two halves  $\mathcal{D}_1$  and  $\mathcal{D}_2$ . We use  $\mathcal{D}_1$  to select a set of variables  $S$  where  $k_n = |S| < n$ . We then apply the pairs bootstrap or a Normal approximation on  $\mathcal{D}_2$  to get a confidence sets for  $\beta_S$  and  $\gamma_S$ . The estimate for  $\beta_S$  is the usual least squares estimator computed from  $\mathcal{D}_2$ . For  $\phi_S$  we use the standard confidence interval for the median (with a Bonferroni correction). For  $\rho_S$  we use the standard asymptotic Normal interval. The details are in Figures 1, 2 and 3. We use the following approximations whose accuracy is studied in Section 3. Surprisingly, the accuracy of the Normal approximation in the case where the linear model is wrong and the dimension increases does not seem to have been studied before.

Let  $\hat{\gamma}_S = m^{-1} \sum_i \delta_i$ ,

$$\delta_i(j) = |Y_i - \hat{\beta}_{S(j),j}^T X_i| - |Y_i - \hat{\beta}_S^T X_i|. \quad (16)$$

Also let

$$v_j^2 = m^{-1} \sum_i (\delta_i(j) - \hat{\gamma}_S(j))^2. \quad (17)$$

We will see that, even with  $k_n$  increasing, the distribution of these parameters can be approximated by Gaussians although the approximation for  $\hat{\beta}$  is poor unless  $k_n$  is small. The accuracy of the Normal approximation is studied in Section 3. (Note: as we shall see in the next Section, we actually add a small amount of random noise to  $\delta_i(j)$  for technical reasons.)



**The Bootstrap.** Let

$$F_n(t) = P^n\left(\sqrt{n}\|\hat{\theta}_S - \theta_S\|_\infty \leq t \mid \mathcal{D}_1\right).$$

We use the bootstrap on  $\mathcal{D}_2$  to estimate  $F_n$ . Let  $P_m$  be the empirical distribution based on the data in  $\mathcal{D}_2$  where  $m = n/2$ . Let  $Z_1^*, \dots, Z_m^*$  denote a bootstrap sample and let  $\hat{\theta}_S^*$  denote the corresponding estimator. Define

$$\hat{F}_n(t) = P^n\left(\sqrt{n}\|\hat{\theta}_S^* - \hat{\theta}_S\|_\infty \leq t \mid Z_1, \dots, Z_n\right).$$

Later we show that, for the Normal and the bootstrap,

$$P^n(\theta \in C_n | \mathcal{D}_1) = 1 - \alpha - O(\Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3})$$

where the errors  $\Delta_{n,1}, \Delta_{n,2}, \Delta_{n,3}$  are deterministic and are uniform over  $\mathcal{P}_n$ . It then follows that, unconditionally,

$$\inf_{P \in \mathcal{P}_n} P^n(\theta \in C_n) = 1 - \alpha - O(\Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3}).$$

It is important that we use the pairs bootstrap — where each pair  $Z_i = (X_i, Y_i)$  is treated as one observation — rather than a residual based bootstrap. The residual bootstrap is only valid if the assumed model is correct. See [Buja et al. \(2015\)](#) for more discussion on this point. For the Normal version we use the Normal approximation in place of the bootstrap. In both cases, the Berry-Esseen theorem for simple convex sets (polyhedra with a limited number of faces) with increasing dimension due to [Chernozhukov et al. \(2013, 2014\)](#) justifies the method. In the case of  $\beta_S$  we also need a Taylor approximation followed by an application of the Gaussian anti-concentration result from the same reference.

For  $\beta_S$ , the estimator is the least squares estimator based on  $\mathcal{D}_2$ . For  $\gamma_S$  and  $\phi_S$ , the estimator  $\hat{\beta}_S$  is any estimator of  $\beta_S$  and is obtained from  $\mathcal{D}_1$ . This can be the lasso estimator, forward stepwise, etc.

**Median.** Recall that  $\phi_S = (\phi_S(j) : j \in S)$  where  $\phi_S(j)$  is the median of  $F_j$  and  $F_j(t) = P^n(W(j) \leq t \mid \mathcal{D}_1)$  is the law of  $\delta(j) = |Y - \hat{\beta}_{S(j),j}^T X_S| - |Y - \hat{\beta}_S^T X_S|$  conditional on  $\mathcal{D}_1$ . Finite sample inference is straightforward using the standard confidence interval for the median based on order statistics. For  $(X_i, Y_i) \in \mathcal{D}_2$  let  $\delta_i(j) = |Y_i - \hat{\beta}_{S(j),j}^T X_i| - |Y_i - \hat{\beta}_S^T X_i|$ , let  $\delta_{(1)}(j) \leq \dots \leq \delta_{(m)}(j)$  be the order statistics and let  $E(j) = [\delta_{(m-a_2)}(j), \delta_{(m-a_1+1)}(j)]$  where we recall that  $m = n/2$ , and

$$a_1 = \frac{m}{2} + \sqrt{m \log \left( \frac{2k}{\alpha} \right)}, \quad a_2 = \frac{m}{2} - \sqrt{m \log \left( \frac{2k}{\alpha} \right)}.$$

Let  $E_n = \bigotimes_{j \in S} E(j)$ . Then

$$\inf_{P \in \mathcal{P}} P^n(\phi_S \in E_n) \geq 1 - \alpha$$

where  $\mathcal{P}$  is the set of all distributions. The steps are summarized in [Figure 3](#).

**Future Prediction Error.** The confidence interval for  $\rho_S$  is

$$F_n = \left[ \hat{\rho}_S - z_\alpha s / \sqrt{m}, \hat{\rho}_S + z_\alpha s / \sqrt{m} \right]$$

where  $\hat{\rho}_S = \frac{1}{m} \sum_{i \in \mathcal{D}_2} \sum_i A_i$ ,  $s^2 = \frac{1}{m} \sum_{i \in \mathcal{D}_2} (A_i - \hat{\rho}_S)^2$  and  $A_i = |Y_i - \hat{\beta}_S^T X_{i,S}|$ .

---

### Boot-Split

INPUT: Data  $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ . Confidence parameter  $\alpha$ . Constant  $\epsilon$  (Section 3.3).

OUTPUT: Confidence sets  $C_n$  and  $D_n$ .

1. Randomly split the data into two halves  $\mathcal{D}_1$  and  $\mathcal{D}_2$ .
2. Use  $\mathcal{D}_1$  to select a subset of variables  $S$ . This can be forward stepwise, the lasso, or any other method. Let  $k = |S|$ .
3. Write  $\mathcal{D}_2 = \{(X_1, Y_1), \dots, (X_m, Y_m)\}$  where  $m = n/2$ . Let  $P_m$  be the empirical distribution of  $\mathcal{D}_2$ .
4. For  $\beta_S$ :
  - (a) Get  $\hat{\beta}_S$  from  $\mathcal{D}_2$  by least squares.
    - i. Draw  $(X_1^*, Y_1^*), \dots, (X_m^*, Y_m^*) \sim P_m$ . Let  $\hat{\beta}_S^*$  be the estimator constructed from the bootstrap sample.
    - ii. Repeat  $B$  times to get  $\hat{\beta}_{S,1}^*, \dots, \hat{\beta}_{S,B}^*$ .
    - iii. Define  $t_\alpha$  by

$$\frac{1}{B} \sum_{b=1}^B I\left(\sqrt{m} \|\hat{\beta}_{S,b}^* - \hat{\beta}_S\|_\infty > t_\alpha\right) = \alpha.$$

- (b) Output:  $C_n = \{\beta \in \mathbb{R}^k : \|\beta - \hat{\beta}_S\|_\infty \leq t_\alpha / \sqrt{m}\}$ .
5. For  $\gamma_S$ :
  - (a) Get  $\hat{\gamma}_S$  from  $\mathcal{D}_1$ . This can be any estimator. For  $j \in S$  let  $\hat{\gamma}_S(j) = \frac{1}{m} \sum_{i=1}^m r_i$  where  $r_i = (\delta_i(j) + \epsilon \xi_i(j))$ ,  $\delta_i(j) = |Y_i - \hat{\beta}_{S,j}^T X_i| - |Y_i - \hat{\beta}_S^T X_i|$  and  $\xi_i(j) \sim \text{Unif}(-1, 1)$ . Let  $\hat{\gamma}_S = (\hat{\gamma}_S(j) : j \in S)$ .
    - i. Draw  $(X_1^*, Y_1^*), \dots, (X_m^*, Y_m^*) \sim P_m$ .
    - ii. Let  $\hat{\gamma}_S(j) = \frac{1}{m} \sum_{i=1}^m r_i^*$ . Let  $\hat{\gamma}_S^* = (\hat{\gamma}_S^*(j) : j \in S)$ .
    - iii. Repeat  $B$  times to get  $\hat{\gamma}_{S,1}^*, \dots, \hat{\gamma}_{S,B}^*$ .
    - iv. Define  $u_\alpha$  by

$$\frac{1}{B} \sum_{b=1}^B I\left(\sqrt{m} \|\hat{\gamma}_{S,b}^* - \hat{\gamma}_S\|_\infty > u_\alpha\right) = \alpha.$$

- (b) Output:  $D_n = \{\gamma_S \in \mathbb{R}^k : \|\gamma_S - \hat{\gamma}_S\|_\infty \leq u_\alpha / \sqrt{m}\}$ .

Figure 1: The Boot-Split Algorithm.

---

---

### Normal-Split

INPUT: Data  $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ . Confidence parameter  $\alpha$ .

OUTPUT: Confidence sets  $C_n$  and  $D_n$ .

1. Randomly split the data into two halves  $\mathcal{D}_1$  and  $\mathcal{D}_2$ .
2. Use  $\mathcal{D}_1$  to select a subset of variables  $S$ . This can be forward stepwise, the lasso, or any other method. Let  $k = |S|$ .
3. For  $\beta_S$ :
  - (a) Get  $\hat{\beta}_S$  from  $\mathcal{D}_2$  by least squares.
  - (b) Output  $C_n = \bigotimes_{j \in S} C(j)$  where  $C(j) = \hat{\beta}_S(j) \pm z_{\alpha/(2k)} \sqrt{\hat{\Gamma}_n(j, j)}$  where  $\hat{\Gamma}$  is given by (41).
4. For  $\gamma_S$ :
  - (a) Get  $\hat{\beta}_S$  from  $\mathcal{D}_1$ . This can be any estimator. For  $j \in S$  let  $\hat{\gamma}_S(j) = \frac{1}{m} \sum_{i=1}^m r_i$  where where  $r_i = (\delta_i(j) + \epsilon \xi_i(j))$ ,  $\delta_i(j) = |Y_i - \hat{\beta}_{S,j}^T X_i| - |Y_i - \hat{\beta}_S^T X_i|$  and  $\xi_i(j) \sim \text{Unif}(-1, 1)$ . Let  $\hat{\gamma}_S = (\hat{\gamma}_S(j) : j \in S)$ .
  - (b) Output  $D_n = \bigotimes_{j \in S} D(j)$  where  $D(j) = \hat{\gamma}_S(j) \pm z_{\alpha/(2k)} v_j$  where  $\hat{V}$  is given by (17).

Figure 2: The Normal-Split Algorithm.

---



---

### Median-Split

INPUT: Data  $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ . Confidence parameter  $\alpha$ .

OUTPUT: Confidence set  $E_n$ .

1. Randomly split the data into two halves  $\mathcal{D}_1$  and  $\mathcal{D}_2$ .
2. Use  $\mathcal{D}_1$  to select a subset of variables  $S$ . This can be forward stepwise, the lasso, or any other method. Let  $k = |S|$ .
3. Write  $\mathcal{D}_2 = \{(X_1, Y_1), \dots, (X_m, Y_m)\}$  where  $m = n/2$ . For  $(X_i, Y_i) \in \mathcal{D}_2$  let

$$W_i(j) = |Y_i - \hat{\beta}_{S,j}^T X_i| - |Y_i - \hat{\beta}_S^T X_i|,$$

4. Let  $W_{(1)}(j) \leq \dots \leq W_{(m)}(j)$  be the order statistics and let  $E(j) = [W_{(m-k_2)}(j), W_{(m-k_1+1)}(j)]$  where

$$k_1 = \frac{m}{2} + \sqrt{m \log \left( \frac{2k}{\alpha} \right)}, \quad k_2 = \frac{m}{2} - \sqrt{m \log \left( \frac{2k}{\alpha} \right)}.$$

5. Let  $E_n = \bigotimes_{j \in S} E(j)$ .

Figure 3: The Median-Split Algorithm.

---

### 3 Berry-Esseen Bounds

The results in the rest of the paper will depend on a Berry-Esseen bound for regression with possibly increasing dimension. In this section, there is no model selection or splitting. We set  $d = k$  and  $S = \{1, \dots, k\}$  where  $k = k_n$  satisfies  $k_n < n$ . Later, these results will be applied after model selection and sample splitting. Existing Berry-Esseen results for nonlinear parameters are given in [Pinelis and Molzon \(2009\)](#); [Shao et al. \(2016\)](#); [Chen and Shao \(2007\)](#); [Anastasiou and Reinert \(2014\)](#); [Anastasiou and Ley \(2015\)](#); [Anastasiou and Gaunt \(2016\)](#). Our results are in the same spirit but we keep careful track of the effect of dimension and the eigenvalues of  $\Sigma$ , while leveraging results from [Chernozhukov et al. \(2013, 2014\)](#) on high dimensional central limit theorems for simple convex sets.

#### 3.1 Berry-Esseen Bounds for Nonlinear Parameters With Increasing Dimension

We start with a general result on the accuracy of the Normal approximation over rectangles for nonlinear parameters. We make use of three results from [Chernozhukov et al. \(2014, 2015\)](#): the high-dimensional central limit theorem, the Gaussian comparison theorem and the Gaussian anti-concentration theorem. In the appendix, we re-state these results in a slightly different form than they appear in the original papers. We do this because we need to keep track of certain constants that affect our results.

Let  $W_1, \dots, W_n \stackrel{i.i.d.}{\sim} P$  where  $W_i \in \mathbb{R}^b$  and  $P \in \mathcal{P}_n$ . Here,  $\mathcal{P}_n$  is a set of distributions that can change with  $n$ . We assume that each  $P \in \mathcal{P}_n$  has a density with respect to Lebesgue measure and that each  $P$  is supported on  $[-A, A]^b$  for some  $0 < A < \infty$ . Let  $g = (g_1, \dots, g_s)^T$  be a vector functions mapping  $\mathbb{R}^b$  to  $\mathbb{R}^s$ . Let  $\theta = g(\psi)$  and  $\hat{\theta} = g(\hat{\psi})$ , where  $\psi = \psi(P) = \mathbb{E}[W_1]$  and  $\hat{\psi} = \frac{1}{n} \sum_{i=1}^n W_i$ . Note that  $\psi \in [-A, A]^b$ . We assume that  $g$  is twice continuously differentiable. For any point  $\psi$  and  $j \in \{1, \dots, s\}$ , we will write  $G_j(\psi)^\top \in \mathbb{R}^b$  and  $H_j(\psi) \in \mathbb{R}^{b \times b}$  for the gradient and Hessian of  $g_j$  at  $\psi$ , respectively. We will set  $G(\psi)$  to be the  $s \times b$  matrix whose  $j^{\text{th}}$  row is  $G_j(\psi)$ . We need the following definitions:

$$\begin{aligned}
 V &= V(P) = \mathbb{E}[(W_i - \psi)(W_i - \psi)^T] & \Gamma &= \Gamma(\psi) = G(\psi)V G(\psi)^T \\
 \underline{\sigma}^2 &= \inf_{P \in \mathcal{P}_n} \min_{j=1, \dots, b} G_j(\psi) V G_j(\psi)^\top & v &= \inf_{P \in \mathcal{P}_n} \lambda_{\min}(V(P)) \\
 \overline{H} &= \sup_{\psi \in [-A, A]^b} \max_{j=1, \dots, s} \lambda_{\max}(H_j(\psi)) & B &= \sup_{\psi \in [-A, A]^b} \max_{j=1, \dots, s} \|G_j(\psi)\| \\
 \aleph_n &= \sqrt{\frac{b \overline{H}^2 B^2 (\log b + \log n)}{n}} & \hat{V} &= \frac{1}{n} \sum_{i=1}^n W_i W_i^\top - \hat{\psi} \hat{\psi}^\top.
 \end{aligned} \tag{18}$$

**Theorem 1.** Assume that  $W_1, \dots, W_n \stackrel{i.i.d.}{\sim} P \in \mathcal{P}_n$ . Let  $Z_n \sim N(0, \Gamma)$ . Then,

$$\sup_{P \in \mathcal{P}_n} \sup_t \left| P^n(\sqrt{n} \|\hat{\theta} - \theta\|_\infty \leq t) - P^n(\|Z_n\|_\infty \leq t) \right| \preceq \Delta_{n,1} + \Delta_{n,2} \tag{19}$$

where

$$\Delta_{n,1} = \max \left\{ \frac{1}{n^{1/6}}, \frac{1}{\sqrt{v}} \left( \frac{b^3(\log 2bn)^7}{n} \right)^{1/6} \right\} \quad (20)$$

$$\Delta_{n,2} = \frac{1}{\underline{\sigma}} \sqrt{\frac{b^2 \bar{H}^2 (\log n)^2 \log b}{n}}. \quad (21)$$

**Remark:** The assumption that the support of  $P$  is compact can be dropped and replaced by moment conditions. The coordinates of the  $W_i$ 's need not be independent. The term  $\Delta_{n,1}$  is the Berry-Esseen error. The term  $\Delta_{n,2}$  is the error due to the non-linearity of the function  $g$ .

*Proof.* For ease of readability we will write  $G_j$  and  $G$  instead of  $G_j(\psi)$  and  $G(\psi)$ , respectively. For each  $j \in \{1, \dots, s\}$ , we use an exact second order Taylor expansion to write

$$\hat{\theta}_j = \theta_j + G_j^T(\hat{\psi} - \psi) + \frac{1}{2n} \delta^T \Lambda_j \delta$$

where  $\delta = \sqrt{n}(\hat{\psi} - \psi)$  and  $\Lambda_j = \int_0^1 H_j((1-u)\psi + u\hat{\psi}) du \in \mathbb{R}^{b \times b}$ . Hence,

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}(\hat{\nu} - \nu) + R \quad (22)$$

where  $\nu = G\psi$ ,  $\hat{\nu} = G\hat{\psi}$  and  $R$  is a random vector in  $\mathbb{R}^s$  whose  $j$ th coordinate is

$$R_j = \frac{1}{2\sqrt{n}} \delta^T \left[ \int_0^1 H_j((1-u)\psi + u\hat{\psi}) du \right] \delta.$$

From Lemma 2 below, we have that exists a constant  $C > 0$ , depending on  $A$  only, such that

$$\sup_{P \in \mathcal{P}} \sup_t \left| P^n(\sqrt{n} \|\hat{\nu} - \nu\|_\infty \leq t) - P^n(\|Z_n\|_\infty \leq t) \right| \leq C \frac{1}{\sqrt{v}} \left( \frac{b^3(\log 2bn)^7}{n} \right)^{1/6}, \quad (23)$$

where  $Z_n \sim N_s(0, \Gamma)$ .

Now we bound the effect of remainder  $R$  in (22). First, since  $\lambda_{\max}(H_j((1-u)\psi + u\hat{\psi})) \leq \bar{H}$ , it is easy to see that

$$\|R\|_\infty \leq \frac{\bar{H} \|\delta\|^2}{2\sqrt{n}},$$

the bound, in fact, holding uniformly in  $P$ . Next, consider the event  $\mathcal{A} = \left\{ \frac{\bar{H} \|\delta\|^2}{2\sqrt{n}} < \epsilon_n \right\}$  where

$$\epsilon_n = \sqrt{\frac{Cb^2 \bar{H}^2 (\log n)^2}{n}}, \quad (24)$$

for a sufficiently large, positive constant  $C$  to be specified later. Then, since the coordinates of  $\widehat{\psi} - \psi$  are sub-Gaussian, and

$$\begin{aligned}
P^n(\mathcal{A}^c) &= P^n\left(\frac{\overline{H}\|\delta\|^2}{2\sqrt{n}} > \epsilon_n\right) = P^n\left(\|\delta\|^2 > \frac{2\sqrt{n}\epsilon_n}{\overline{H}}\right) \\
&\leq P^n\left(b\max_j \delta_j^2 > \frac{2\sqrt{n}\epsilon_n}{\overline{H}}\right) = P^n\left(\max_j \delta_j^2 > \frac{2\sqrt{n}\epsilon_n}{b\overline{H}}\right) \\
&\leq \sum_j P^n\left(\delta_j^2 > \frac{2\sqrt{n}\epsilon_n}{b\overline{H}}\right) = \sum_j P^n\left(|\delta_j| > \sqrt{\frac{2\sqrt{n}\epsilon_n}{b\overline{H}}}\right) \\
&= \sum_j P^n\left(|\widehat{\psi}_j - \psi_j| > \sqrt{\frac{2\epsilon_n}{\sqrt{n}b\overline{H}}}\right) \leq 2b \exp\left(-\frac{\sqrt{n}\epsilon_n}{4A^2b\overline{H}}\right) \leq \left(\frac{1}{n}\right)^{1/6}
\end{aligned}$$

where the second-to-last inequality is Hoeffding inequality for bounded variables and the last inequality follows by taking the constant  $C$  in the definition of  $\epsilon_n$  in (24) appropriately large. In fact, the bound on the probability of the event  $\mathcal{A}^c$  holds uniformly over all  $P \in \mathcal{P}_n$ .

Next, for any  $t > 0$  and uniformly in  $P \in \mathcal{P}$ ,

$$\begin{aligned}
P^n(\sqrt{n}\|\widehat{\theta} - \theta\|_\infty \leq t) &= P^n(\sqrt{n}\|\widehat{\theta} - \theta\|_\infty \leq t, \mathcal{A}) + P^n(\sqrt{n}\|\widehat{\theta} - \theta\|_\infty \leq t, \mathcal{A}^c) \\
&\leq P^n(\sqrt{n}\|\widehat{\nu} - \nu\|_\infty \leq t + \epsilon_n) + P^n(\mathcal{A}^c) \\
&= P^n(\|Z_n\|_\infty \leq t + \epsilon_n) + C \frac{1}{\sqrt{v}} \left(\frac{b^3(\log 2bn)^7}{n}\right)^{1/6} + P^n(\mathcal{A}^c) \quad (25)
\end{aligned}$$

where the inequality follows from (22) and the second identity from the Berry-Esseen bound (23). By the Gaussian anti-concentration inequality of Theorem 26,

$$P^n(\|Z_n\|_\infty \leq t + \epsilon_n) \leq P^n(\|Z_n\|_\infty \leq t) + \frac{\epsilon_n}{\underline{\sigma}}(\sqrt{2\log b} + 2).$$

Using the previous inequality on the first term of (25), we obtain

$$\begin{aligned}
P^n(\sqrt{n}\|\widehat{\theta} - \theta\|_\infty \leq t) &\leq P^n(\|Z_n\|_\infty \leq t) + C \left[ \frac{\epsilon_n}{\underline{\sigma}}(\sqrt{2\log b} + 2) + \frac{1}{\sqrt{v}} \left(\frac{b^3(\log 2bn)^7}{n}\right)^{1/6} \right] + P^n(\mathcal{A}^c) \\
&\leq P^n(\|Z_n\|_\infty \leq t) + C \left[ \frac{\epsilon_n}{\underline{\sigma}}(\sqrt{2\log b} + 2) + \max \left\{ \frac{1}{n^{1/6}}, \frac{1}{\sqrt{v}} \left(\frac{b^3(\log 2bn)^7}{n}\right)^{1/6} \right\} \right],
\end{aligned}$$

where in the second inequality we have used the fact that  $P^n(\mathcal{A}^c) \leq n^{-1/6}$ . By a symmetric argument, we have

$$P^n(\sqrt{n}\|\widehat{\theta} - \theta\|_\infty \leq t) \geq P^n(\|Z_n\|_\infty \leq t) - C \left[ \frac{\epsilon_n}{\underline{\sigma}}(\sqrt{2\log b} + 2) + \max \left\{ \frac{1}{n^{1/6}}, \frac{1}{\sqrt{v}} \left(\frac{b^3(\log 2bn)^7}{n}\right)^{1/6} \right\} \right].$$

The result now follows by (24). ■

The following lemma shows that the linear term  $\sqrt{n}(\hat{\nu} - \nu)$  has a Gaussian-like behavior in a sense made precise by the following result, which is an application of the Berry-Esseen Theorem 27, due to Chernozhukov et al. (2014).

**Lemma 2.** *There exists a constant  $C > 0$ , depending on  $A$  only, such that*

$$\sup_{P \in \mathcal{P}} \sup_t \left| P^n(\sqrt{n} \|\hat{\nu} - \nu\|_\infty \leq t) - P^n(\|Z_n\|_\infty \leq t) \right| \leq C \frac{1}{\sqrt{v}} \left( \frac{b^3(\log 2bn)^7}{n} \right)^{1/6}, \quad (26)$$

where  $Z_n \sim N_s(0, \Gamma)$ .

*Proof.* Let  $\mathcal{V} = \mathcal{V}(G) = \{v_1, \dots, v_{2s}\}$ , where for  $l = 1, 2, \dots, s$ , we define  $v_{2l-1} = \frac{G_l^\top}{\|G_l\|}$  and  $v_{2l} = -\frac{G_l^\top}{\|G_l\|}$ . For a given  $t > 0$  and for any Jacobian matrix  $G = G(\psi)$ , set

$$P(G, t) = \left\{ x \in \mathbb{R}^b : v_j^\top x \leq t_j, \forall v_j \in \mathcal{V}(G) \right\}, \quad (27)$$

where, for  $l = 1, \dots, 2$ ,  $t_{2l-1} = t_{2l} = \frac{t}{\|G_l\|}$ . Then,

$$\|\sqrt{n}(\hat{\nu} - \nu)\|_\infty \leq t \quad \text{if and only if} \quad \sqrt{n}(\hat{\psi} - \psi) \in P(G, t).$$

Similarly for  $Z_n \sim N_k(0, \Gamma)$  and  $\tilde{Z}_n \sim N_b(0, V)$ ,

$$\|Z_n\|_\infty \leq t \quad \text{if and only if} \quad \tilde{Z}_n \in P(G, t).$$

Now consider the class  $\mathcal{A}$  of all subsets of  $\mathbb{R}^b$  of the form specified in (27), where  $t$  ranges over the positive reals and  $G$  ranges in  $\{G(\psi(P)), P \in \mathcal{P}\}$ . Notice that this class is comprised of polytopes with at most  $2s$  facets. Also, from the discussion above,

$$\sup_{P \in \mathcal{P}_n} \sup_{t > 0} \left| P^n(\|\sqrt{n}(\hat{\nu} - \nu)\|_\infty \leq t) - P^n(\|Z_n\|_\infty \leq t) \right| = \sup_{A \in \mathcal{A}} \left| P^n(\sqrt{n}(\hat{\psi} - \psi) \in A) - P^n(\tilde{Z}_n \in A) \right|. \quad (28)$$

The claimed result follows from applying the Berry-Esseen bound for polyhedral classes, Theorem 27 in the appendix, due to Chernozhukov et al. (2014) to the term on the left hand side of (28). To that end, we first need to ensure that conditions (M1'), (M2') and (E1') in that Theorem are satisfied.

For a given  $G = G(\psi)$  and all  $i = 1, \dots, n$ , set  $\tilde{W}_i = (\tilde{W}_{i,1}, \dots, \tilde{W}_{i,2s}) = (W_i^\top v, v \in \mathcal{V}(G))$ . To see that condition (M1') holds, notice that, for each  $j = 1, \dots, 2s$ ,

$$\mathbb{E} \left[ \tilde{W}_{i,j}^2 \right] \geq \min_j v_j^\top V v_j \geq \lambda_{\min}(V),$$

where  $V = \text{Cov}[W]$ . To verify condition (M2'), notice that, for for each  $j = 1, \dots, 2s$  and  $k = 1, 2$ , we have that

$$\mathbb{E} \left[ |\tilde{W}_{i,j}|^{2+k} \right] \leq \mathbb{E} \left[ \|W_i\|^{2+k} \right]$$

Now, since  $\|W_i\| \leq A\sqrt{b}$ , we see that by setting  $B_n = A^4 b^{3/2}$  condition (M2') is satisfied. Finally, condition (E1') is easily satisfied, possibly by increasing the constant in the term  $B_n$ .

Thus, Theorem 27 gives

$$\sup_{A \in \mathcal{A}} \left| P^n(\sqrt{n}(\hat{\psi} - \psi) \in A) - P^n(\tilde{Z}_n \in A) \right| \leq C \frac{1}{\sqrt{\lambda_{\min}(V)}} \left( \frac{b^3 (\log 2bn)^7}{n} \right)^{1/6},$$

and the result follows from (28) and the fact that  $\lambda_{\min}(V) \geq v$  for all  $P \in \mathcal{P}_n$ .  $\blacksquare$

To use the Normal approximation for inference, we need to estimate the variance matrix  $\Gamma = G(\psi)V(\psi)G(\psi)^T$ . In conventional fixed-dimension asymptotics, we could appeal to Slutsky's theorem and ignore the effect of replacing  $\Gamma$  with an estimate. But in computing Berry-Esseen bounds with increasing dimension we cannot ignore the effect of estimating  $\Gamma$ . First, we need to bound the elementwise difference between  $\Gamma$  and  $\hat{\Gamma} = G(\hat{\psi})\hat{V}G(\hat{\psi})^T$ .

**Lemma 3.** *There exists a constant  $C$  dependent on  $A$  such that*

$$\sup_{P \in \mathcal{P}_n} P^n \left( \max_{j,l} \left| \hat{\Gamma}(j,l) - \Gamma(j,l) \right| \geq C \aleph_n \right) \leq \frac{1}{n} \quad (29)$$

where we recall that  $\aleph_n$  and  $\overline{H}$  are defined in (18).

*Proof.* Write  $\hat{\Gamma}_{st} = \hat{\Gamma}(s,t)$  and  $\Gamma_{st} = \Gamma(s,t)$ . Note that  $\hat{\Gamma} - \Gamma$  can be written as

$$\begin{aligned} & (\hat{G} - G)VG^T + GV(\hat{G} - G)^T + (\hat{G} - G)V(\hat{G} - G)^T + \\ & (\hat{G} - G)(\hat{V} - V)G^T + G(\hat{V} - V)(\hat{G} - G)^T + G(\hat{V} - V)G^T + (\hat{G} - G)(\hat{V} - V)^T(\hat{G} - G)^T. \end{aligned}$$

The first and second terms are dominant, so it will be enough to compute a high-probability bound for  $(\hat{G} - G)VG^T$ . For any  $j$  and  $l \in \{1, \dots, s\}$ , we have

$$\left| (\hat{G}_j - G_j) V G_l^T \right| \leq \lambda_{\max}(V) \|\hat{G}_j - G_j\| \|G_l\| \leq A^2 B \max_j \|\hat{G}_j - G_j\|, \quad (30)$$

where the second inequality stems from the definition of  $B$  and the fact that the largest eigenvalue of (the positive definite matrix)  $V$  is bounded by  $A^2$ , since all of its entries are. It remains to bound the stochastic term  $\max_j \|\hat{G}_j - G_j\|$ . Towards that end, we will show that, for some constant  $C$  dependent on  $A$  only,

$$P^n \left( \max_j \|\hat{G}_j - G_j\| \leq C \overline{H} \sqrt{b(\log b + \log n)/n} \right) \geq 1 - 1/n. \quad (31)$$

Indeed, by a Taylor expansion,

$$\hat{G}_j - G_j = (\hat{\psi} - \psi)^T \int_0^1 H_j((1-u)\psi + u\hat{\psi}) du \quad \text{for all } j \in \{1, \dots, s\},$$



so that

$$\max_j \|\widehat{G}_j - G_j\| \leq \|\psi - \widehat{\psi}\| \max_j \left\| \int_0^1 H_j((1-u)\psi + u\widehat{\psi}) du \right\|_{\text{op}}.$$

Since the coordinates of  $\widehat{\psi}$  are bounded in absolute value by  $A$ , and using the inequality  $\|\widehat{\psi} - \psi\| \leq 2A\sqrt{b}\|\widehat{\psi} - \psi\|_\infty$ , an application of Hoeffding inequality and of the union bound yield that, for some positive constant  $C$  dependent on  $A$  only,  $P^n\left(\|\widehat{\psi} - \psi\| \leq \sqrt{b(\log b + \log n)/n}\right) \geq 1 - 1/n$ , for all  $P \in \mathcal{P}_n^\dagger$ . We restrict to this event. By convexity of operator norm  $\|\cdot\|_{\text{op}}$  and our assumption, we have that

$$\max_j \left\| \int_0^1 H_j((1-u)\psi + u\psi) du \right\|_{\text{op}} \leq \overline{H},$$

yielding the bound in (31). Combined with (30), we conclude that on an event of probability at least  $1 - 1/n$ ,  $\max_{j,l} |\widehat{\Gamma}(j,l) - \Gamma(j,l)| \preceq \aleph_n$ .  $\blacksquare$

Now we construct the confidence set. Let  $Q = (Q(1), \dots, Q(b))$  be i.i.d. standard Normal variables, independent of the data. Let  $\widetilde{Z} = \widehat{\Gamma}^{1/2}Q$  and define  $t_\alpha$  by

$$P^n(\|\widetilde{Z}\|_\infty > t_\alpha | \widehat{\Gamma}) = \alpha. \quad (32)$$

Finally, let

$$R_n = \left\{ \theta \in \mathbb{R}^s : \|\theta - \widehat{\theta}\|_\infty \leq \frac{t_\alpha}{\sqrt{n}} \right\}. \quad (33)$$

**Theorem 4.** *We have that*

$$\inf_{P \in \mathcal{P}} P^n(\theta \in R_n) = 1 - \alpha - O\left(\Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3} + \frac{1}{n}\right), \quad (34)$$

where

$$\Delta_{n,3} = C \frac{\aleph_n^{1/3} (2 \log 2s)^{2/3}}{\underline{\sigma}^{1/3}}. \quad (35)$$

*Proof.* Let  $Z_n \sim N(0, \Gamma)$ . We have

$$P^n(\theta \notin R_n) = P^n(\sqrt{n}\|\widehat{\theta} - \theta\|_\infty > t_\alpha) = P^n(\|\widetilde{Z}_n\|_\infty > t_\alpha) + a_1(t) + a_2(t)$$

where

$$a_1(t) = P^n(\sqrt{n}\|\widehat{\theta} - \theta\|_\infty > t_\alpha) - P^n(\|Z_n\|_\infty > t_\alpha)$$

and

$$a_2(t) = P^n(\|Z_n\|_\infty > t) - P^n(\|\widetilde{Z}_n\|_\infty > t_\alpha).$$

Now

$$P^n(\|\widetilde{Z}_n\|_\infty > t_\alpha) \stackrel{d}{=} P^n(\|\widehat{\Gamma}^{1/2}Q\|_\infty > t_\alpha) = \mathbb{E}[P^n(\|\widehat{\Gamma}^{1/2}Q\|_\infty > t_\alpha | \widehat{\Gamma})] = \alpha.$$

Theorem (1) implies that  $\sup_t |a_1(t)| \preceq \Delta_n$ . To bound  $a_2$ , let  $\mathcal{E}_n = \{\max_{j,k} |\hat{\Gamma} - \Gamma| \leq C\aleph_n\}$ . Then,  $P^n(\mathcal{E}_n) \geq 1 - 1/n$ , uniformly over all  $P$  in  $\mathcal{P}_n$ . Next, we have that

$$a_2(t) \leq \mathbb{E}[P^n(\|\Gamma^{1/2}Q\|_\infty > t) - P^n(\|\hat{\Gamma}^{1/2}Q\|_\infty > t_\alpha|\hat{\Gamma}); \mathcal{E}] + P^n(\mathcal{E}_n^c),$$

where  $\mathbb{E}[\cdot; \mathcal{E}_n]$  denotes expectation restricted to the event  $\mathcal{E}_n$ . By the Gaussian comparison Theorem 28 the term inside the expected value is bounded by  $\Delta_{n,3}$ . ■

Instead of  $L_\infty$  balls, we can also construct our confidence set to be a hyper-rectangle, with side lengths proportional to the standard errors of the projection parameters. That is, we define

$$\tilde{C}_n = \bigotimes_{j \in S} C(j), \quad (36)$$

where  $C(j) = \hat{\beta}_S(j) \pm z_{\alpha/(2s)} \sqrt{\hat{\Gamma}_n(j, j)}$  with  $\hat{\Gamma}$  given by (41) and  $z_{\alpha/(2s)}$  the upper  $1 - \alpha/(2s)$  quantile of a standard normal variate. Notice that we use a Bonferroni correction to guarantee a nominal coverage of  $1 - \alpha$ . Also, note that  $z_{\alpha/(2s)} = O(\sqrt{\log s})$ .

**Theorem 5.** *We have that*

$$\inf_{P \in \mathcal{P}_n} P^n(\theta \in \tilde{C}_n) \geq (1 - \alpha) - \frac{1}{n} - \frac{\aleph_n z_{\alpha/(2s)}}{(\min_j \gamma_j)^2} \left( \sqrt{2 + \log(2s)} + 2 \right) - \Delta_{n,1} - \Delta_{n,2}.$$

*Proof.* For  $j = 1, \dots, s$ , let  $\gamma_j = \sqrt{\Gamma_{j,j}}$ ,  $\hat{\gamma}_j = \sqrt{\hat{\Gamma}_{j,j}}$  and  $\hat{t}_j = z_{\alpha/(2s)} \hat{\gamma}_j$ . We use the same arguments and notation as in the proof of Theorem 1 and of Lemma 2. Thus, let  $\mathcal{A}$  be the event that  $\frac{\bar{H} \|\delta\|^2}{2\sqrt{n}} < \epsilon_n$ , where  $\frac{\bar{H} \|\delta\|^2}{2\sqrt{n}}$  is an upper bound on  $\|R\|_\infty$ , with  $R$  the reminder in the Taylor series expansion (22) and  $\epsilon_n$  as in (24).

Next, for each  $t \in \mathbb{R}_+^{2s}$  and any Jacobian matrix  $G = G(\psi(P))$ , with  $P \in \mathcal{P}$ , let

$$P(G, t) = \left\{ x \in \mathbb{R}^b : v_j^\top x \leq t_j, \forall v_j \in \mathcal{V}(G) \right\}, \quad (37)$$

where  $\mathcal{V}(G)$  is defined in Lemma 2. Then, for any positive numbers  $(t'_1, \dots, t'_2)$

$$|\sqrt{n}(\hat{\nu}_j - \nu_j)|_\infty \leq t'_j, j = 1, \dots, s \quad \text{if and only if} \quad \sqrt{n}(\hat{\psi} - \psi) \in P(G, t),$$

where the coordinates of  $t \in \mathbb{R}^{2s}$  are as follows: for  $l = 1, \dots, s$ ,  $t_{2l-1} = t_{2l} = \frac{t'_j}{\|G_j\|}$ .

Consider now the class of subsets of  $\mathbb{R}^b$  of the form specified in (37), where  $t$  ranges over the positive vectors in  $\mathbb{R}^{2s}$  and  $G$  ranges in  $\{G(\psi(P)), P \in \mathcal{P}\}$ . This is a class comprised by polytopes with at most  $2s$  faces in  $\mathbb{R}^b$ . Thus, using the same arguments as in the proof of Lemma 2, we obtain that

$$\sup_{t=(t_1, \dots, t_s)^\top \in \mathbb{R}_+^s} \left| P^n \left( \sqrt{n}|\hat{\theta}_j - \theta_j| \leq t_j, \forall j \right) - P^n(|Z_{n,j}| \leq t_j, \forall j) \right| \leq C \frac{1}{\sqrt{v}} \left( \frac{b^3 (\log 2bn)^7}{n} \right)^{1/6},$$

for some  $C > 0$  depending only on  $A$ . Using this result, we can conclude that

$$\begin{aligned}
P^n(\sqrt{n}|\hat{\theta}_j - \theta_j| \leq \hat{t}_j, \forall j) &= P^n(\sqrt{n}|\hat{\theta}_j - \theta_j| \leq \hat{t}_j, \forall j; \mathcal{A}) + P^n(\sqrt{n}|\hat{\theta}_j - \theta_j| \leq \hat{t}_j, \forall j; \mathcal{A}^c) \\
&\leq P^n(\sqrt{n}|\hat{\nu}_j - \nu_j| \leq \hat{t}_j + \epsilon_n, \forall j) + P^n(\mathcal{A}^c) \\
&= P^n(|Z_{n,j}| \leq \hat{t}_j + \epsilon_n, \forall j) + C \frac{1}{\sqrt{v}} \left( \frac{b^3(\log 2bn)^7}{n} \right)^{1/6} + P^n(\mathcal{A}^c) \\
&\leq P^n(|Z_{n,j}| \leq \hat{t}_j, \forall j) + C \left[ \frac{\epsilon_n}{\underline{\sigma}} (\sqrt{2 \log b} + 2) + \frac{1}{\sqrt{v}} \left( \frac{b^3(\log 2bn)^7}{n} \right)^{1/6} + P^n(\mathcal{A}^c) \right],
\end{aligned}$$

where the last inequality follows from Nazarov's inequality, see Lemma A1. in [Chernozhukov et al. \(2014\)](#). A similar argument gives

$$P^n(\sqrt{n}|\hat{\theta}_j - \theta_j| \leq \hat{t}_j, \forall j) \geq P^n(|Z_{n,j}| \leq \hat{t}_j, \forall j) - C \left[ \frac{\epsilon_n}{\underline{\sigma}} (\sqrt{2 \log b} + 2) + \frac{1}{\sqrt{v}} \left( \frac{b^3(\log 2bn)^7}{n} \right)^{1/6} - P^n(\mathcal{A}^c) \right].$$

To conclude the proof, it will be enough to show that

$$P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \hat{\gamma}_j, \forall j) \geq (1 - \alpha) - \frac{1}{n} - \frac{\Xi_n z_{\alpha/(2s)}}{(\min_j \gamma_j)^2} \left( \sqrt{2 + \log(2s)} + 2 \right).$$

Towards that end, let  $\mathcal{E}$  be the event that  $\max |\gamma_j - \hat{\gamma}_j| \leq \Xi_n$  and assume that  $P^n(\mathcal{E}^c) \leq 1/n$ . Then,

$$\begin{aligned}
P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \hat{\gamma}_j, \forall j) &= P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \hat{\gamma}_j, \forall j) - P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \gamma_j, \forall j) + P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \gamma_j, \forall j) \\
&\geq P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \hat{\gamma}_j, \forall j; \mathcal{E}) - P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \gamma_j, \forall j) + P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \gamma_j, \forall j) \\
&\geq P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \hat{\gamma}_j, \forall j; \mathcal{E}) - P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \gamma_j, \forall j) + (1 - \alpha).
\end{aligned}$$

Next,

$$P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \hat{\gamma}_j, \forall j; \mathcal{E}) \geq P^n(|Z_{n,j}| \leq z_{\alpha/(2s)}(\gamma_j - \Xi_n), \forall j; \mathcal{E}) \geq P^n(|Z_{n,j}| \leq z_{\alpha/(2s)}(\gamma_j - \Xi_n), \forall j) - P^n(\mathcal{E}^c).$$

Thus,

$$\begin{aligned}
P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \hat{\gamma}_j, \forall j) &\geq (1 - \alpha) - P^n(\mathcal{E}^c) + P^n(|Z_{n,j}| \leq z_{\alpha/(2s)}(\gamma_j - \Xi_n), \forall j) - P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \gamma_j, \forall j) \\
&\geq (1 - \alpha) - \frac{1}{n} - \frac{\Xi_n z_{\alpha/(2s)}}{\min_j \gamma_j} \left( \sqrt{2 + \log(2s)} + 2 \right).
\end{aligned}$$

since  $P^n(\mathcal{E}^c) \leq \frac{1}{n}$  and, by Nazarov's inequality,

$$P^n(|Z_{n,j}| \leq z_{\alpha/(2s)}(\gamma_j - \Xi_n), \forall j) - P^n(|Z_{n,j}| \leq z_{\alpha/(2s)} \gamma_j, \forall j) \geq \frac{\Xi_n z_{\alpha/(2s)}}{\min_j \gamma_j} \left( \sqrt{2 + \log(2s)} + 2 \right).$$

We now claim that  $\Xi_n = \frac{\aleph_n}{\min_j \gamma_j}$ . Indeed

$$|\hat{\gamma}_j - \gamma_j| = \frac{|\hat{\gamma}_j - \gamma_j| |\hat{\gamma}_j + \gamma_j|}{|\hat{\gamma}_j + \gamma_j|} = \frac{|\hat{\gamma}_j^2 - \gamma_j^2|}{|\hat{\gamma}_j + \gamma_j|} \leq \frac{|\hat{\gamma}_j^2 - \gamma_j^2|}{\gamma_j} \leq \frac{\aleph_n}{\min_j \gamma_j},$$

and the bound  $P^n(\mathcal{E}^c) \leq \frac{1}{n}$  follows from Lemma 3. ■

## The Bootstrap

A similar result holds for the bootstrap. We first need a preliminary, standard result.

**Lemma 6.** *There exists positive a constant  $C$  depending on  $A$  such that,*

$$\sup_{P \in \mathcal{P}_n} P^n(\|V - \hat{V}\|_{\text{op}} > C\mathfrak{T}_n) < \frac{1}{n}$$

where

$$\mathfrak{T}_n = b \min \left\{ \sqrt{\frac{b + \log 2n}{n}}, \frac{b + \log(2n)}{n} \right\}. \quad (38)$$

*Proof.* The vectors  $W_i$ 's are sub-Gaussian vectors with variance parameter  $\tau^2 = bA^2$ . This is because, for any  $v \in \mathbb{R}^b$  of unit norm,  $|v^\top X| \leq \sqrt{b}A$ . The claims then follows from well-known finite sample bounds for the operator norm of  $V - \hat{V}$ ; see [Vershynin \(2010\)](#). ■

**Theorem 7.** *Assume the same conditions and notation as in Theorem 1. Let*

$$\begin{aligned} F_n(t) &= P^n(\sqrt{n}|\hat{\theta} - \theta|_\infty \leq t) \\ \hat{F}_n(t) &= P^n(\sqrt{n}|\hat{\theta}^* - \hat{\theta}|_\infty \leq t \mid Z_1, \dots, Z_n). \end{aligned}$$

*Assume in addition that  $n$  is large enough so that  $\sigma_n^2 = \underline{\sigma}^2 - C\aleph_n > 0$  and and  $v_n := v - \mathfrak{T}_n > 0$  where  $C$  is the larger of the two constants in Lemma 3 and Lemma 6. Then, there is an event  $\mathcal{E}_n$  such that  $\sup_{P \in \mathcal{P}^\dagger} P^n(\mathcal{E}_n^c) \leq 2/n$  and, on  $\mathcal{E}_n$ ,*

$$\sup_{P \in \mathcal{P}^\dagger} \sup_t \left| \hat{F}_n(t) - F_n(t) \right| \leq \tilde{\Delta}_{n,1} + \tilde{\Delta}_{n,2} + \Delta_{n,3}, \quad (39)$$

where  $\Delta_{n,3}$  is given by (35) and

$$\begin{aligned} \tilde{\Delta}_{n,1} &= \max \left\{ \frac{1}{n^{1/6}}, \frac{1}{\sqrt{v}} \left( \frac{b^3(\log 2bn)^7}{n} \right)^{1/6} \right\} \\ \tilde{\Delta}_{n,2} &= \frac{1}{\underline{\sigma}_n} \sqrt{\frac{b^2 \bar{H}^2 (\log n)^2 \log b}{n}}. \end{aligned}$$

Let  $C_n = \{\theta : \|\theta - \hat{\theta}\|_\infty \leq t_\alpha/\sqrt{n}\}$  where  $t_\alpha = \hat{F}_n^{-1}(1 - \alpha)$ . Then

$$\inf_{P \in \mathcal{P}_n} P^n(\theta \in C_n) = 1 - \alpha - O \left( \tilde{\Delta}_{n,1} + \tilde{\Delta}_{n,2} + \Delta_{n,3} + \frac{1}{n} \right). \quad (40)$$

The result in Theorem 5 also holds for the bootstrap.

*Proof.* Let  $Z_n \sim N(0, \Gamma)$  where  $\Gamma = GVG^T$ . Take  $\mathcal{E}_n$  to be the event that

$$\left\{ \max_{j,k} |\hat{\Gamma} - \Gamma| \leq C\aleph_n \right\} \cap \left\{ \|V - \hat{V}\|_{\text{op}} \leq C\daleth_n \right\}$$

where  $C$  is the larger of the two constants in Lemma 3 and Lemma 6. Then the probability of  $\mathcal{E}_n$  is at least  $1 - 2/n$ , uniformly over all the distributions in  $\mathcal{P}_n^\dagger$ . Let  $\tilde{Z}_n \sim N(0, \hat{\Gamma})$  where  $\hat{\Gamma} = \hat{G}\hat{V}\hat{G}^T$  and  $\hat{G} = G(\hat{\psi})$ ,  $\hat{V} = n^{-1} \sum_{i=1}^n (W_i - \hat{\psi})(W_i - \hat{\psi})^T$ . Then

$$\sup_t |\hat{F}_n(t) - F_n(t)| \leq a_1 + a_2 + a_3$$

where

$$\begin{aligned} a_1 &= \sup_t |\hat{F}_n(t) - \mathbb{P}(\|\tilde{Z}_n\|_\infty \leq t)| \\ a_2 &= \sup_t |\mathbb{P}(\|\tilde{Z}_n\|_\infty \leq t) - \mathbb{P}(\|Z_n\|_\infty \leq t)| \\ a_3 &= \sup_t |\mathbb{P}(\|Z_n\|_\infty \leq t) - F_n(t)|. \end{aligned}$$

Now  $a_3$  has already been bounded in the earlier Berry-Esseen theorem. For  $a_2$  we use the Gaussian comparison Theorem 28 as in the proof of Theorem 4 to conclude that  $a_2 \preceq \Delta_{n,3}$ . To bound  $a_1$ , we use essentially the same argument used in Theorem 1. Simply replace  $\psi$  with  $\hat{\psi}$  and replace  $\hat{\psi}$  with  $\hat{\psi}^*$ . The assumption that  $n$  is large enough so that  $\underline{\sigma}_n > 0$  and  $v_n > 0$  are required in order to ensure that, with probability at least  $1 - 2/n$ ,  $\min_j \sqrt{\hat{\Gamma}(j, j)} > \sqrt{\underline{\sigma}^2 - C\aleph_n} > 0$  and the minimal eigenvalue of  $\hat{V}$  is no smaller than  $v - C\daleth_n > 0$ . Thus, the same arguments used in the proof of Theorem 1 but restricted to the event  $\mathcal{E}_n$  yield that, on this event,

$$\sup_{P \in \mathcal{P}_n} \sup_t \left| \hat{F}_n(t) - P^n(\|\tilde{Z}_n\|_\infty \leq t) \right| \preceq \tilde{\Delta}_{n,1} + \tilde{\Delta}_{n,2}.$$

Finally, (40) follows from the continuity of  $F_n$ . The proof of the result in Theorem 5 for the bootstrap is similar and so is omitted.  $\blacksquare$

### 3.2 Berry-Esseen Bounds for the Projection Parameter

We now apply the results derived in the previous section to vector of projection parameters defined in (10), assuming a data generating distribution from the class  $\mathcal{P}_n^\dagger$  defined in (5). Recall that, for a random vector  $(X, Y) \in \mathbb{R}^{k+1}$  generated from some  $P \in \mathcal{P}^\dagger$ , the vector of projection parameters is given by  $\beta = \Sigma^{-1}\alpha \in \mathbb{R}^k$ , where  $\Sigma = \mathbb{E}[XX^\top] \in \mathbb{R}^{k \times k}$  and  $\alpha = \mathbb{E}[YX] \in \mathbb{R}^k$ . We can represent the parameters  $(\Sigma, \alpha)$  in vectorized form as

$$\psi = \psi(P) = \begin{bmatrix} \text{vech}(\Sigma) \\ \alpha \end{bmatrix} \in \mathbb{R}^b,$$

where  $b = \frac{k^2+3k}{2} = O(k^2)$  and, for a symmetric  $k \times k$  matrix  $A$ ,  $\text{vech}(A)$  denotes the column vector of dimension  $k(k+1)/2$  obtained by vectorizing only the lower triangular part of  $A$ . Given an i.i.d. sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  from  $P$ , we set

$$W_i = \begin{bmatrix} \text{vech}(X_i X_i^\top) \\ Y_i \cdot X_i \end{bmatrix} \in \mathbb{R}^b, \quad i = 1, \dots, n$$

and  $\widehat{\psi} = \frac{1}{n} \sum_{i=1}^n W_i$ . Note that  $\mathbb{E}[\widehat{\psi}] = \psi$ .

Then, we can express both the projection parameter  $\beta$  and the least square estimator  $\widehat{\beta} = \widehat{\Sigma}^{-1} \widehat{\alpha}$  where  $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$  and  $\widehat{\alpha} = \frac{1}{n} \sum_{i=1}^n Y_i X_i$ , as non-linear functions of  $\psi$  and  $\widehat{\psi}$ , respectively. (Note that  $\widehat{\Sigma}$  is invertible with probability one.) That is, we may write

$$\beta = g(\psi) \quad \text{and} \quad \widehat{\beta} = g(\widehat{\psi}),$$

respectively, where  $g: \mathbb{R}^b \rightarrow \mathbb{R}^k$  is given by

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto (\text{math}(x_1))^{-1} x_2,$$

where  $x_1$  and  $x_2$  correspond to the first  $k(k+1)/2$  and the last  $k$  coordinates of  $x$ , respectively, and  $\text{math}$  is the inverse mapping of  $\text{vech}$ , i.e.  $\text{math}(x) = A$  if and only if  $\text{vech}(A) = x$ .

Furthermore, recall that

$$V = \text{Var}[W_i] = \mathbb{E}[(W_i - \psi)(W_i - \psi)^T] \quad \text{and} \quad \widehat{V} = \frac{1}{n} \sum_{i=1}^n [(W_i - \widehat{\psi})(W_i - \widehat{\psi})^T],$$

and

$$\Gamma = G(\psi) V G(\psi)^T \quad \text{and} \quad \widehat{\Gamma} = G(\widehat{\psi}) \widehat{V} G(\widehat{\psi})^T \quad (41)$$

where  $G = G(\psi)$  is given in (48). ( $\Gamma$  is the variance of the linear approximation to  $\widehat{\beta}$ .)

To bound the approximation error in regression — where  $\widehat{\beta}$  is the least squares estimator — we need to bound  $\underline{\sigma}$ ,  $\overline{H}$  and  $B$  and then apply the theorems in the previous subsection.

**Lemma 8.** *Let  $\beta(j) = e_j^\top \beta \equiv g_j(\psi)$  where  $e_j$  is a vector whose  $j^{\text{th}}$  element is 1 and is 0 otherwise. The gradient and Hessian of  $g_j$  are given by*

$$G_j = e_j^\top \left( \left[ -(\alpha^\top \otimes I_k) (\Omega \otimes \Omega) \quad \Omega \right] \right) D_h \quad (42)$$

and

$$H_j = D_h^\top A_j D_h, \quad (43)$$

respectively, where

$$A_j = \frac{1}{2} \left( (I_b \otimes e_j^\top) H + H^\top (I_b \otimes e_j) \right),$$

and

$$H = \begin{bmatrix} -((\Omega \otimes \Omega) \otimes I_k) \begin{bmatrix} 0_{k^3 \times k^2} & (I_k \otimes \text{vec}(I_k)) \end{bmatrix} + (I_{k^2} \otimes (\alpha^\top \otimes I_k)) G \begin{bmatrix} (\Omega \otimes \Omega) & 0_{k^2 \times k} \end{bmatrix} \\ \begin{bmatrix} -(\Omega \otimes \Omega) & 0_{k^2 \times k} \end{bmatrix} \end{bmatrix},$$

and  $D_h$  is the modified duplication matrix defined by  $D\psi_h = \psi$ , with  $\psi_h$  the vector consisting of the subset of  $\psi$  not including entries that correspond to the upper diagonal entries of  $\Sigma$ . Assume that  $k \geq u^2$ . Then

$$B = \max_j \sup_{\psi} \|G_j(\psi)\| \preceq \frac{\sqrt{k}}{u^2}, \quad \bar{H} = \max_j \sup_{\psi} \lambda_{\max}(H_j(\psi)) \preceq \frac{k}{u^3}, \quad (44)$$

and

$$\underline{\sigma} = \inf_{P \in \mathcal{P}_n} \min_j \sqrt{G_j V G_j^\top} \geq \frac{\sqrt{v}}{U}. \quad (45)$$

**Remark.** The assumption that  $k \geq u^2$  is not actually needed but this is the most common case and it simplifies the expressions a bit.

*Proof.* Recall that  $\psi = (\sigma^\top, \alpha^\top)^\top$ , where  $\sigma = \text{vec}(\Sigma)$ , and  $\xi = (w^\top, \alpha^\top)^\top$  with  $w = \text{vec}(\Omega)$ . For  $1 \leq j \leq n$ , let

$$\beta_j = g_j(\psi) = e_j^\top \Omega \alpha,$$

where  $e_j$  is the  $j$ th elements of the standard basis in  $\mathbb{R}^n$ . Then, we can write

$$g_j(\psi) = g(f(\psi)),$$

where  $f(\psi) = \xi$  and  $g(\xi) = e_j^\top \Omega \alpha$ . Notice that  $f: \mathbb{R}^{k(k+1)} \rightarrow \mathbb{R}^{k(k+1)}$  and  $g: \mathbb{R}^{k(k+1)} \rightarrow \mathbb{R}$ . The dimension of  $\psi$  and  $\xi$  is  $b = k^2 + k$ .

Using the chain rule, the gradient of  $g_j(\psi)$  is

$$G_j = Dg_j(\psi) = Dg(\xi)Df(\psi) = e_j^\top \left[ \left( \alpha^\top \otimes I_k \right) E + \Omega F \right] \begin{bmatrix} -\Omega \otimes \Omega & 0 \\ 0 & I_k \end{bmatrix},$$

where

$$E = \begin{bmatrix} I_{k^2} & 0_{k^2 \times k} \end{bmatrix} = \frac{dv}{d\psi}$$

is a  $k^2 \times b$  matrix and

$$F = \begin{bmatrix} 0_{k \times k^2} & I_k \end{bmatrix} = \frac{d\alpha}{d\psi}$$

is  $k \times b$ . Now

$$\begin{aligned} \left( \alpha^\top \otimes I_k \right) E \begin{bmatrix} -\Omega \otimes \Omega & 0 \\ 0 & I_k \end{bmatrix} &= \left( \alpha^\top \otimes I_k \right) \begin{bmatrix} I_{k^2} & 0_{k^2 \times k} \end{bmatrix} \begin{bmatrix} -\Omega \otimes \Omega & 0 \\ 0 & I_k \end{bmatrix} \\ &= \begin{bmatrix} -\left( \alpha^\top \otimes I_k \right) (\Omega \otimes \Omega) & 0_{k \times k} \end{bmatrix} \end{aligned}$$

Next,

$$\begin{aligned} \Omega F \begin{bmatrix} -\Omega \otimes \Omega & 0 \\ 0 & I_k \end{bmatrix} &= \Omega \begin{bmatrix} 0_{k \times k^2} & I_k \end{bmatrix} \begin{bmatrix} -\Omega \otimes \Omega & 0 \\ 0 & I_k \end{bmatrix} \\ &= \Omega \begin{bmatrix} 0_{k \times k^2} & I_k \end{bmatrix} = \begin{bmatrix} 0_{k \times k^2} & \Omega \end{bmatrix}. \end{aligned}$$

Thus, plugging the last two expressions into the initial formula for  $Dg_j(\psi)$  we get

$$\begin{aligned} G_j = Dg_j(\psi) &= e_j^\top \left( \begin{bmatrix} -(\alpha^\top \otimes I_k) (\Omega \otimes \Omega) & 0_{k \times k} \end{bmatrix} + \begin{bmatrix} 0_{k \times k^2} & \Omega \end{bmatrix} \right) \\ &= e_j^\top \left( \begin{bmatrix} -(\alpha^\top \otimes I_k) (\Omega \otimes \Omega) & \Omega \end{bmatrix} \right) \end{aligned} \quad (46)$$

$$(47)$$

Let  $G = d\beta(j)/d\psi$  so that

$$G = \begin{pmatrix} G_1^T \\ \vdots \\ G_k^T \end{pmatrix}. \quad (48)$$

Now, to get  $Hf_j(\psi)$  we need to compute the derivative of  $(Dg_j(\psi))^\top$ . Since  $Dg_j(\psi)$  is a (row) vector,

$$D((Dg_j(\psi))^\top) = D(Dg_j(\psi)).$$

So we proceed with the simpler computation of  $D(Dg_j(\psi))$ , which is

$$(I_b \otimes e_j^\top) \frac{d \operatorname{vec} \left( \begin{bmatrix} -(\alpha^\top \otimes I_k) (\Omega \otimes \Omega) & \Omega \end{bmatrix} \right)}{d\psi}$$

where the first matrix is of dimension  $b \times kb$  and the second matrix is of dimension  $kb \times b$ , so that  $Hf_j(\psi)$  is  $b \times b$ . Then,

$$\frac{d \operatorname{vec} \left( \begin{bmatrix} -(\alpha^\top \otimes I_k) (\Omega \otimes \Omega) & \Omega \end{bmatrix} \right)}{d\psi} = \begin{bmatrix} -\frac{d(\alpha^\top \otimes I_k)(\Omega \otimes \Omega)}{d\psi} \\ \frac{d\Omega}{d\psi} \end{bmatrix}. \quad (49)$$

The derivative at the bottom of the previous expression is:

$$\frac{d\Omega}{d\psi} = \frac{d\Omega}{d\Sigma} \frac{d\Sigma}{d\psi} = -(\Omega \otimes \Omega)E = -(\Omega \otimes \Omega)[I_{k^2} \quad 0_{k^2 \times k}] = \begin{bmatrix} -(\Omega \otimes \Omega) & 0_{k^2 \times k} \end{bmatrix}.$$

The top derivative in (49) is more involved. By the product rule:

$$\frac{d(\alpha^\top \otimes I_k)(\Omega \otimes \Omega)}{d\psi} = \left( (\Omega \otimes \Omega) \otimes I_k \right) \frac{d(\alpha^\top \otimes I_k)}{d\psi} + \left( I_{k^2} \otimes (\alpha^\top \otimes I_k) \right) \frac{d(\Omega \otimes \Omega)}{d\psi}.$$

The first derivative in the last expression is

$$\begin{aligned} \frac{d(\alpha^\top \otimes I_k)}{d\psi} &= \frac{d(\alpha^\top \otimes I_k)}{d\alpha} \frac{d\alpha}{d\psi} = (I_k \otimes T_{1,k} \otimes I_k)(I_k \otimes \operatorname{vec}(I_k))F \\ &= (I_k \otimes \operatorname{vec}(I_k))F = (I_k \otimes \operatorname{vec}(I_k)) \begin{bmatrix} 0_{k \times k^2} & I_k \end{bmatrix} \\ &= \begin{bmatrix} 0_{k^3 \times k^2} & (I_k \otimes \operatorname{vec}(I_k)) \end{bmatrix}, \end{aligned}$$



where the third identity follows since  $T_{k,1} = I_k$  and, therefore,  $(I_k \otimes T_{1,k} \otimes I_k) = I_{k^3}$ .

As for the other derivative,

$$\begin{aligned} \frac{d(\Omega \otimes \Omega)}{d\psi} &= \frac{d(\Omega \otimes \Omega)}{d\Omega} \frac{d\Omega}{d\Sigma} \frac{d\Sigma}{d\psi} = -J(\Omega \otimes \Omega)E \\ &= -J(\Omega \otimes \Omega) \begin{bmatrix} I_{k^2} & 0_{k^2 \times k} \end{bmatrix} = -J \begin{bmatrix} (\Omega \otimes \Omega) & 0_{k^2 \times k} \end{bmatrix}, \end{aligned}$$

where

$$J = \left[ (I_k \otimes \Omega) \otimes I_{k^2} \right] \left( I_k \otimes T_{k,k} \otimes I_k \right) \left( I_{k^2} \otimes \text{vec}(I_k) \right) + \left[ I_{k^2} \otimes (\Omega \otimes I_k) \right] \left( I_k \otimes T_{k,k} \otimes I_k \right) \left( \text{vec}(I_k) \otimes I_{k^2} \right).$$

To see this, notice that, by the product rule we have

$$J = \frac{d(\Omega \otimes \Omega)}{d\Omega} = \frac{d(\Omega \otimes I_k)(I_k \otimes \Omega)}{d\Omega} = \left[ (I_k \otimes \Omega) \otimes I_{k^2} \right] \frac{d(\Omega \otimes I_k)}{d\Omega} + \left[ I_{k^2} \otimes (\Omega \otimes I_k) \right] \frac{d(I_k \otimes \Omega)}{d\Omega}.$$

Next,

$$\frac{d(\Omega \otimes I_k)}{d\Omega} = \left( I_k \otimes T_{k,k} \otimes I_k \right) \left( I_{k^2} \otimes \text{vec}(I_k) \right) = \left( I_{k^2} \otimes T_{k,k} \right) \left( I_k \otimes \text{vec}(I_k) \otimes I_k \right)$$

and

$$\frac{d(I_k \otimes \Omega)}{d\Omega} = \left( I_k \otimes T_{k,k} \otimes I_k \right) \left( \text{vec}(I_k) \otimes I_{k^2} \right) = \left( T_{k,k} \otimes I_{k^2} \right) \left( I_k \otimes \text{vec}(I_k) \otimes I_k \right).$$

The formula for  $J$  follows from the last three expressions. Notice that  $J$  is matrix of size  $k^4 \times k^2$ . Finally, plugging the expressions for  $\frac{d(\alpha^\top \otimes I_k)(\Omega \otimes \Omega)}{d\psi}$  and  $\frac{d\Omega}{d\psi}$  in (49) we get that the Hessian  $Hf_j(\psi)$  is

$$\frac{1}{2} \left( (I_b \otimes e_j^\top) H + H^\top (I_b \otimes e_j) \right)$$

where

$$H = \begin{bmatrix} - \left( (\Omega \otimes \Omega) \otimes I_k \right) \begin{bmatrix} 0_{k^3 \times k^2} & (I_k \otimes \text{vec}(I_k)) \end{bmatrix} + \left( I_{k^2} \otimes (\alpha^\top \otimes I_k) \right) J \begin{bmatrix} (\Omega \otimes \Omega) & 0_{k^2 \times k} \end{bmatrix} \\ \begin{bmatrix} -(\Omega \otimes \Omega) & 0_{k^2 \times k} \end{bmatrix} \end{bmatrix}.$$

So far we have ignored the fact that  $\Sigma$  is symmetric. To account for the symmetry, the Hessian is

$$D_h^\top Hf_j(\psi) D_h,$$

where  $D_h$  is the modified duplication matrix such that  $D\psi_h = \psi$  where  $\psi_h$  is the vector comprised by the sub-vector of  $\psi$  not including the entries corresponding to the upper (or lower) diagonal entries of  $\Sigma$ .

Now we bound these terms. Recall that  $B = \sup_{\psi \in C} \max_j \sum_s |G_j(s)|$ . We see in equation (46), that each element of  $G_j$  is bounded above by a constant times

$$\lambda_{\max}(\Omega \otimes \Omega) = (\lambda_{\max}(\Omega))^2 \leq \frac{1}{u^2}.$$

So

$$\|G_j\| \leq \|e_j\|((\|\alpha\| \times \|\Omega\|^2) + \|\Omega\|) \preceq \frac{\sqrt{k}}{u^2} + \frac{1}{u} \preceq \frac{\sqrt{k}}{u^2},$$

since  $k \geq u^2$ , giving the first bound on  $B$  in (44).

Turning to the Hessian, we bound the largest eigenvalue of the Hessian by bounding largest singular value of the individual terms in the above expression. Let  $\sigma_1$  denote the largest singular value. Then

$$\sigma_1([\Omega \otimes \Omega, 0]) = \sigma_1(\Omega \otimes \Omega) = \lambda_{\max}^2(\Omega) = 1/u^2.$$

Next,

$$\sigma_1((\Omega \otimes \Omega \otimes I_k)(0, I_k \otimes \text{vec}(I_k))) = \sigma_1([0, \Omega \times \text{vec}(\Omega)]) = \sigma_1(\Omega \times \text{vec}(\Omega)).$$

But

$$\sigma_1(\text{vec}(\Omega)) = \|\Omega\|_F = \sqrt{\sum_{i=1}^k \sigma_i(\Omega)} \leq \sqrt{k} \sigma_1(\Omega) = \frac{\sqrt{k}}{u^2}.$$

Next we bound the term  $[I_{k^2} \otimes \alpha^T \otimes I_k]J[(\Omega \otimes \Omega)0_{k^2 \times k}]$ . Now

$$\begin{aligned} \sigma_1(J) &\leq 2\sigma_1((I_k \otimes \Omega \otimes I_{k^2})(I_k \otimes T_{k,k} \otimes I_k)(I_{k^2} \otimes \text{vec}(I_k))) \\ &= 2\sigma_1(\Omega)\|I_k\|_F = 2\sqrt{k}\sigma_1(\Omega) = \frac{2\sqrt{k}}{u}. \end{aligned}$$

Also,  $\sigma_1([I_{k^2} \otimes \alpha^T \otimes I_k]) \leq \|\alpha\|$ . Hence,

$$\sigma_1[I_{k^2} \otimes \alpha^T \otimes I_k]J[(\Omega \otimes \Omega)0_{k^2 \times k}] \leq 2\sqrt{k}\|\alpha\|\sigma_1^3(\Omega) \preceq \frac{k}{u^3}.$$

We thus have

$$\sigma_1(H) \leq \sigma_1^2(\Omega) + \sqrt{k}\sigma_1^2(\Omega) + \frac{k}{u^3} \leq \frac{\sqrt{k}}{u^2} + \frac{k}{u^3}.$$

Finally,

$$\begin{aligned} \sigma_1(H_j) &= \sigma_1\left(\frac{1}{2}((I_b \otimes e_j)H + H^T(I_b \otimes e_j))\right) \leq \sigma_1((I_b \otimes e_j)H) \leq \sigma_1(I_b)\sigma_1(e_j)\sigma_1(H) \\ &\leq \sigma_1^2(\Omega) + \sqrt{k}\sigma_1^2(\Omega) + 2\sqrt{k}\|\alpha\|\sigma_1^2(\Omega) \preceq \frac{\sqrt{k}}{u^2} + \frac{k}{u^3} \leq \frac{k}{u^3} \end{aligned}$$

whenever  $u \leq \sqrt{k}$ . This gives the bound on  $\overline{H}$  in (44). Finally, the bound on  $\underline{\sigma}$  given in (45) follows from (42). Indeed, for every  $P \in \mathcal{P}^\dagger$

$$\min_j \sqrt{G_j V G_j^\top} \geq \sqrt{v} \min_j \|G_j\|.$$

Then, using (42),

$$\min_j \|G_j\| \geq \min_j \|\Omega_j\| \geq \lambda_{\min}(\Omega) = \frac{1}{U},$$

where  $\Omega_j$  denotes the  $j$ -th row of  $\Omega$ . The result follows. ■

Now we can insert these bounds into Theorem 4 and, noting that  $b = O(k^2)$ , we immediately get the following.

**Theorem 9.** *Let  $R_n$  be the confidence rectangle defined in (33). Then*

$$\inf_{P \in \mathcal{P}_n^\dagger} P^n(\beta \in R_n) = 1 - \alpha - O(\Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3}) \quad (50)$$

where

$$\begin{aligned} \Delta_{n,1} &= \max \left\{ \frac{1}{n^{1/6}}, \frac{1}{\sqrt{v}} \left( \frac{k^6 (\log kn)^7}{n} \right)^{1/6} \right\} \\ \Delta_{n,2} &= \frac{U}{\sqrt{v}} \sqrt{\frac{k^6 \log^2 n \log k}{n u^6}} \\ \Delta_{n,3} &= \left( \frac{U}{\sqrt{v}} \right)^{1/3} \left( \frac{k^5 \log k + \log n}{u^{10} n} \log^4 k \right)^{1/6}. \end{aligned}$$

The same holds for the bootstrap confidence set, with the terms  $\frac{U}{\sqrt{v}}$  replaced by

$$\left( \frac{\sqrt{v}}{U} - \sqrt{\frac{k^5}{u^{10} \frac{\log k + \log n}{n}}} \right)^{-1},$$

for all  $n$  for which the above term is positive and also with  $v$  replaced by  $v - \mathfrak{T}_n$  where

$$\mathfrak{T}_n = k^2 \min \left\{ \sqrt{\frac{k^2 + \log 2n}{n}}, \frac{k^2 + \log(2n)}{n} \right\}.$$

*Proof.* Follows from Lemma 8 together with Theorem 4 and Lemma 6. ■

### 3.3 Bounds for LOCO Parameters

Now we turn to the parameter  $\gamma$ . Let  $w, w_j \in \mathbb{R}^{k_n}$  be two fixed vectors. Define  $\gamma = (\gamma(1), \dots, \gamma(k_n))$  where

$$\gamma(j) = \mathbb{E}_{X,Y} \left[ |Y - w_j^T X| - |Y - w^T X| \right]$$

where  $(X, Y)$  is a new pair. Later, when we do sample splitting,  $w$  will be replaced by the estimator from the first half of the data and  $w_j$  will be an estimator obtained after removing covariate  $j$  and re-running the model selection on the first half of the data. For now,  $w$  and  $w_j$  are just two fixed vectors and we assume that  $\|w\|_\infty \leq C < \infty$  and  $\max_j \|w_j\|_\infty < C < \infty$ . The estimate of  $\gamma$  is

$$\hat{\gamma}(j) = \frac{1}{n} \sum_{i=1}^n \left[ |Y_i - w_j^T X_i| - |Y_i - w^T X_i| \right] \equiv \frac{1}{n} \sum_{i=1}^n \delta_i(j).$$

To derive a CLT for  $\hat{\gamma}(j)$  we face a technical problem. If  $k$  is large, it is possible that any one variable could have a tiny influence on the predictions. This means that we could have  $\text{Var}(\delta_i(j)) \rightarrow 0$  as  $k \rightarrow \infty$ . This, in general, precludes a central limit theorem. To address this, we re-define  $\hat{\gamma}(j)$  by

$$\hat{\gamma}(j) = \frac{1}{n} \sum_{i=1}^n (\delta_i(j) + \epsilon \xi_j) \quad (51)$$

where  $\epsilon > 0$  is an arbitrarily small positive constant and  $\xi_j \sim \text{Uniform}(-1, 1)$ . This prevents the variance from vanishing. Adding this extra noise has the effect of making the inference conservative: the confidence intervals will be slightly wider. For any non-trivial value of  $\gamma(j)$  this has a negligible effect. It is possible to let  $\epsilon \rightarrow 0$  as  $n \rightarrow \infty$  at the expense of a slower Berry-Esseen rate. For simplicity, we take  $\epsilon$  to be a fixed, small constant.

Note that if a variable is not selected then  $\gamma(j)$  is identically 0. Obviously these parameter estimates will not be asymptotically Normal. We restrict to  $j \in \mathcal{J}$ , the set of variables that are selected in the model selection phase.

The next Theorem is the CLT for  $\hat{\gamma}$ . Since  $\hat{\gamma}$  is now just a vector of averages with non-vanishing variances, the accuracy is higher than in the previous Theorem. Also, there is no need for a Taylor approximation so the Theorem is almost an immediate consequence of Theorem 2.2 of [Chernozhukov et al. \(2014\)](#).

**Theorem 10.** *Let  $Z_n \sim N(0, S_n)$  where  $S_n = \mathbb{E}[(\delta_i - \gamma)(\delta_i - \gamma)^T]$ . Then,*

$$\sup_t \left| P^n(\sqrt{n} \|\hat{\gamma} - \gamma\|_\infty \leq t) - P^n(\|Z\|_\infty \leq t) \right| \leq \frac{1}{\epsilon} \left( \frac{(\log k_n)^7}{n} \right)^{1/6}. \quad (52)$$

Define  $D_n(\alpha) = \left\{ \gamma : \|\gamma - \hat{\gamma}\|_\infty \leq \frac{t_\alpha}{\sqrt{n}} \right\}$  where  $t_\alpha = \hat{F}_n^{-1}(1 - \alpha)$  and

$$\hat{F}_n(t) = P^n \left( \sqrt{n} \|\hat{\gamma}^*(b) - \hat{\gamma}(b)\|_\infty \leq t_\alpha \mid Z_1, \dots, Z_n \right) = 1 - \alpha.$$

As the statistic is linear, the next result follows from Theorem K.1 of [Chernozhukov et al. \(2014\)](#) and their Gaussian comparison theorem.

**Theorem 11.** *We have*

$$\sup_{P \in \mathcal{P}_n} \left| P^n(\gamma \in D_n(\alpha)) - (1 - \alpha) \right| \leq \frac{1}{\epsilon} \left( \frac{(\log(nk_n))^7}{n} \right)^{1/6} \left( \log \left( \frac{nk_n}{\log(nk_n)} \right) \right)^{1/3}. \quad (53)$$

A similar result holds for the Normal approximation. As the proof is almost the same, we omit it.

The validity of the coverage of  $E_n$  is a trivial consequence of the union bound and the fact that  $E_n(j)$  is the usual confidence interval for the median. This we have:

**Proposition 12.** *For every  $n$ ,*

$$\inf_{P \in \mathcal{P}_n} P^n(\phi_S \in E_n) \geq 1 - \alpha. \quad (54)$$

Finally, we have from the usual one-dimensional Berry-Esseen theorem:

**Proposition 13.** *For every  $n$ ,*

$$\inf_{P \in \mathcal{P}_n} P^n(\rho_S \in F_n) \geq 1 - \alpha - O(1/\sqrt{n}). \quad (55)$$

### 3.4 The Sparse Case

Now we briefly discuss the case of sparse fitting where  $k_n \leq k = O(1)$  so that the size of the selected model is not allowed to increase with  $n$ . In this case, things simplify considerably. The standard central limit theorem shows that

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow N(0, \Gamma)$$

where  $\Gamma = \Sigma^{-1} \mathbb{E}[(Y - \beta^T X)^2] \Sigma^{-1}$ . Furthermore,  $\Gamma$  can be consistently estimated by the sandwich estimator  $\hat{\Gamma} = \hat{\Sigma}^{-1} A \hat{\Sigma}^{-1}$  where  $A = n^{-1} \mathbb{X}^T R \mathbb{X}$ ,  $\mathbb{X}_{ij} = X_i(j)$ ,  $R$  is a the  $k \times k$  diagonal matrix with  $R_{ii} = (Y_i - X_i^T \hat{\beta})^2$ . By Slutsky's theorem, valid asymptotic confidence sets can be based on the Normal distribution with  $\hat{\Gamma}$  in place of  $\Gamma$  (Buja et al. (2015)).

However, if  $k$  is non-trivial relative to  $n$ , then fixed  $k$  asymptotics may be misleading. In this case, the results of the previous section may be more appropriate. In particular, replacing  $\Gamma$  with an estimate then has a non-trivial effect on the coverage accuracy. Furthermore, the accuracy depends on  $1/u$  where  $u = \lambda_{\min}(\Sigma)$ . But when we apply the results after sample splitting (as is our goal), we need to define  $u$  as  $u = \min_{|S| \leq k} \lambda_{\min}(\Sigma_S)$ . As  $d$  increases,  $u$  can get smaller and smaller even with fixed  $k$ . Hence, the usual fixed  $k$  asymptotics may be misleading.

## 4 Accuracy of Inferences From Sample Splitting

Now we return to sample splitting. A model  $S$  is selected from  $\mathcal{D}_1$  where  $k_n = |S| < n$ . We then construct the confidence sets described in the previous section using the data from  $\mathcal{D}_2$ .

From the results in the last section we conclude that  $P^n(\theta_S \in C_S | \mathcal{D}_1) = 1 - \alpha + O(\Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3})$  where the error terms do not depend on the data. Hence,

$$P^n(\theta_S \in C_S) = \mathbb{E}[P^n(\theta_S \in C_S | \mathcal{D}_1)] = 1 - \alpha + O(\Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3})$$

and so all the coverage accuracy statements of the previous section apply unconditionally. However, the definition of  $u$  is now

$$u = \min_{|S| \leq k} \lambda_{\min}(\Sigma_S). \quad (56)$$

Recall that for the LOCO parameter estimate, we add a small amount of noise as in (51). To keep  $\delta_i(j)$  bounded, we also truncate  $\hat{\beta}_S(j)$  so that  $|\hat{\beta}_S(j)| \leq L$  for some  $L$ .

It remains to bound  $\|\hat{\beta}_S - \beta_S\|_\infty$ ,  $\|\hat{\gamma}_S - \gamma_S\|_\infty$ ,  $\nu(C_n)$  and  $\nu(D_n)$ .

**Theorem 14.** *Suppose that  $|S| \leq k_n$  where  $k_n < n$ . Then, with probability tending to one, uniformly over  $\mathcal{P}$ ,*

$$\|\hat{\beta}_S - \beta_S\|_\infty \preceq \frac{k_n^{3/2}}{u^2} \sqrt{\frac{\log k + \log n}{n}} \quad (57)$$

and

$$\|\hat{\gamma}_S - \gamma_S\|_\infty \preceq \sqrt{\frac{\log k_n}{n}}. \quad (58)$$

*Proof.* We prove the first bound; the second follows similarly. We have

$$\begin{aligned} \hat{\beta}_S(j) - \beta_S(j) &= g_j(\hat{\psi}) - g_j(\psi) = (\hat{\psi}_S - \psi_S)^T \int_0^1 G_j((1-u)\psi + u\hat{\psi}_S) du \\ &\leq \|\hat{\psi}_S - \psi_S\| \left\| \int_0^1 G_j((1-u)\psi + u\hat{\psi}_S) du \right\| \preceq k \sqrt{\frac{\log n + \log k}{n}} \frac{\sqrt{k}}{u^2}, \end{aligned}$$

where we have used the fact, established in the proof of Lemma 3, that  $\|\hat{\psi}_S - \psi\| \preceq k \sqrt{\frac{\log n + \log k}{n}}$  with probability at least  $1/n$ , the convexity of the Euclidean norm and the bound on  $\max_j \|G_j\|$  in (44). Hence,  $|\hat{\beta}_S(j) - \beta_S(j)| \preceq \frac{k_n^{3/2}}{u^2} \sqrt{(\log k + \log n)/n}$ . ■

For the size of the confidence sets we have the following.

**Theorem 15.** *For both  $\beta_S$  and  $\gamma_S$  we have*

$$\sup_{P \in \mathcal{P}_n} P^n \left( \nu(R_n) > \sqrt{\frac{C \log k_n}{n}} \right) \rightarrow 0.$$

*Proof.* This follows from the fact that the max of  $k$  Normals is of order  $\sqrt{\log k}$ . ■

Now consider the confidence rectangle  $E_n$  for  $\phi_S$ .

**Theorem 16.** *Suppose that the distribution of each  $\delta_i(j)$  is continuous with non-zero density in a neighborhood of its median. Then each side of the rectangle  $E_n$  has width  $O_P(\sqrt{\log k/n})$ .*

*Proof.* We can write  $E(j) = [\delta_{(a)}, \delta_{(b)}]$  where

$$a = \frac{m}{2} - \sqrt{m \log \left( \frac{2k}{\alpha} \right)}, \quad b = \frac{m}{2} + \sqrt{m \log \left( \frac{2k}{\alpha} \right)}.$$

Let  $F_m$  be the empirical distribution of the  $\delta_i(j)$ 's. Then we can write

$$a = mF_m^{-1}(\beta_1), \quad b = nF_m^{-1}(\beta_2)$$

where

$$\beta_1 = \frac{1}{2} - \sqrt{\frac{1}{m} \log \left( \frac{2k}{\alpha} \right)}, \quad \beta_2 = \frac{1}{2} + \sqrt{\frac{1}{m} \log \left( \frac{2k}{\alpha} \right)}.$$

Let  $p$  be the density of  $\delta_i(j)$  and let  $F$  be the cdf. From the Bahadur representation of the sample quantiles (see Chapter 6 of [Csorgo \(1983\)](#)) we have, for any  $\beta$ , that

$$F_n^{-1}(\beta) = \theta + \frac{\beta - F_n(\theta)}{p(\theta)} + R_n$$

where  $\theta = F^{-1}(\beta)$  and

$$\frac{|p(\theta)R_n(\beta)|n^{3/4}}{\sqrt{\log m} (\log \log m)^{1/4}} = O_P(1).$$

The result follows. ■

## 5 Improving the Coverage Accuracy of the Bootstrap

The coverage accuracy for  $\gamma_S$  and  $\phi_S$  is very high. But the inferences for  $\beta_S$  are less accurate if  $k_n$  is allowed to increase with  $n$ . Of course, one way to ensure accurate inferences is simply to focus on  $\gamma_S$  or  $\phi_S$  instead of  $\beta_S$ . Here we discuss some other approaches to ensure coverage accuracy.

If we use ridge regression instead of least squares, the gradient and Hessian of  $\beta$  are bounded and the error terms are very small. However, this could degrade prediction accuracy. This leads to a tradeoff between inferential accuracy and prediction accuracy. Investigating this tradeoff will be left to future work.

Some authors have suggested the estimator  $\hat{\beta} = \tilde{\Sigma}^{-1}\hat{\alpha}$  where  $\tilde{\Sigma}$  is a block diagonal estimator of  $\Sigma$ . If we restrict the block size to be bounded above by a constant, then we get back the accuracy of the sparse regime. Again there is a tradeoff between inferential accuracy and prediction accuracy.

The accuracy of the bootstrap can be increased by using the *image bootstrap* as we now describe. First we apply the bootstrap to get a confidence set for  $\psi$ . Let

$$H_n = \left\{ \psi_S : \|\psi_S - \hat{\psi}_S\|_\infty \leq \frac{t_\alpha}{\sqrt{n}} \right\} \quad (59)$$

where  $t_\alpha$  is the bootstrap quantile defined by  $\widehat{F}(t_\alpha) = 1 - \alpha$  and

$$\widehat{F}(t) = P(\sqrt{n}||\widehat{\psi}_S^* - \widehat{\psi}_S||_\infty \leq t \mid Z_1, \dots, Z_n). \quad (60)$$

Since  $\psi$  is just a vector of moments, it follows from Theorem K.1 of [Chernozhukov et al. \(2014\)](#) that

$$\sup_{P \in \mathcal{P}_n} |P^n(\psi \in H_n) - (1 - \alpha)| \leq \frac{1}{a} \left( \frac{(\log k_n)^7}{n} \right)^{1/6}$$

and the dimension enters only logarithmically. (Recall that  $a$  is defined in (18). Recall that  $\beta_S = g(\psi_S)$ . Now define

$$C_n = \left\{ g(\psi) : \psi \in H_n \right\}. \quad (61)$$

We call  $C_n$  the *image bootstrap confidence set* as it is just the nonlinear function  $g$  applied to the confidence set  $H_n$ . The image bootstrap is usually avoided because it generally leads to conservative confidence sets. But, as we see here, in the high-dimensional case, the gain in coverage accuracy can be dramatic.

**Theorem 17.** *We have that*

$$\inf_{P \in \mathcal{P}'_n} P^n(\beta \in C_n) \geq 1 - \alpha - \frac{1}{a} \left( \frac{\log k_n}{n} \right)^{1/6}. \quad (62)$$

Also,

$$\sup_{P \in \mathcal{P}_n^\dagger} \mathbb{E}[\nu(C_n)] \leq \sqrt{\frac{k^3 \log k}{n u^2}}. \quad (63)$$

Hence, the diameter tends uniformly to 0 if  $k_n(\log k_n)^{1/3} = o(n^{1/3})$ . Interestingly, this is the same condition required in [Portnoy \(1987\)](#) although the setting is quite different.

*Proof.* Since  $\psi \in H_n$  implies that  $\beta = g(\psi) \in C_n$ , we then have that, uniformly over  $\mathcal{P}_n$ ,

$$P^n(\beta \in C_n) \geq 1 - \alpha - \frac{1}{a} \left( \frac{\log k_n}{n} \right)^{1/6}.$$

To bound the size, let  $\beta \in H_n$  and use the same arguments in the proof of Theorem 14. ■

The implied confidence set for  $\beta(j)$  is

$$C_j = \left[ \inf_{\psi \in H_n} g(\psi), \sup_{\psi \in H_n} g(\psi) \right].$$

Currently, we do not have a computationally efficient method to find the supremum and infimum. A crude approximation is given by taking a random sample  $\psi_1, \dots, \psi_N$  from  $H_n$  and taking

$$a(j) \approx \min_j g(\widehat{\psi}_j), \quad b(j) \approx \max_j g(\widehat{\psi}_j).$$

In future work, we will investigate more efficient methods.



## 6 The Variability Due to Sample Splitting

A concern about sample splitting is that splitting introduces randomness into the inference (although the same could be said about  $K$ -fold cross-validation). If we repeated the procedure we could get different inferences. In section we give some informal discussion about some methods to assess this variability. We will focus on  $\beta_S$ .

### 6.1 Multisplit Confidence Sets

A natural way to assess the variability due to sample splitting is to split many times. Suppose, then, that we repeat the splitting  $N$  times. This leads to  $N$  different model selections  $S_1, \dots, S_t, \dots, S_N$  and corresponding projection parameters  $\beta(S_1), \dots, \beta(S_N)$ .

We could then construct corresponding confidence sets  $C_1, \dots, C_N$ . Is there a way to summarize this variability? One possibility is to somehow combine these confidence sets. This is the approach taken in [Meinshausen et al. \(2012\)](#). They assume that the linear model is correct, that the model is sparse and that high probability, the model chosen from each split contains the true non-zero variables. This ensures that the  $\beta(j)$ 's have the same meaning in each split and so the confidence intervals can be combined. In our assumption-free setting, these assumptions are no longer true. If  $S$  and  $S'$  are two different models, both containing  $j$ , then  $\beta_j(S)$  and  $\beta_j(S')$  have a different meaning.

However, we can get a sense of the variability due to sample splitting by examining the confidence sets for various splits. To this end, we proceed as follows. First, we extend  $\beta_{S_t}$  by setting  $\beta_{S_t}(j) = 0$  if  $j \notin S_t$ . Next, each  $C_t$  is extended by including  $\{0\}$  whenever  $j \notin S_t$ . Specifically, note that  $C_t = \bigotimes_{j \in S_t} [a_{tj}, b_{tj}]$ . For  $j \notin S_t$  we set  $a_{tj} = b_{tj} = 0$  and we set

$$C_t = \bigotimes_{j=1}^d [a_{tj}, b_{tj}] \subset \mathbb{R}^d.$$

Now we define the *meta-confidence set* by

$$C_N^\dagger = \bigcup_{t=1}^N C_t.$$

For any given variable  $j$ , define  $C_N^\dagger(j) = [\min_t a_{tj}, \max_t b_{tj}]$ . Plotting the  $C_N^\dagger(j)$ 's versus  $j$  gives a simple visual summary of the variability of the confidence sets.

The meta-confidence set represents two sources of variability: the variability within each model selection and the variability due to the instability of the model selection procedure. We construct each confidence set at level  $1 - \alpha/N$ . We then have, by the union bound,

$$P^n(\beta_{S_t} \in C_N, \text{ for all } 1 \leq t \leq N) \geq 1 - \alpha \quad (64)$$

which ensures that each confidence set is valid for its parameter. We do not, however, make any attempt to interpret the meta-confidence set as a confidence statement about some overall, fixed parameter.

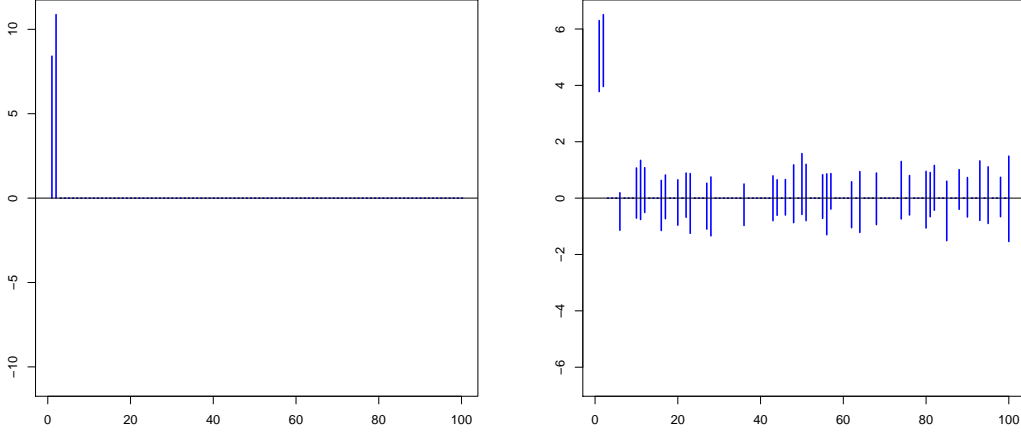


Figure 4: Left: With two strong signals and  $k = 1$  steps of forward stepwise, the meta-confidence interval shows the model selection variability. Right: With  $k = 5$ , the intervals for the first two variables (the signals) are much smaller since the model selection variability is smaller.

**Example 18.** Suppose that  $d = 100$  and that

$$Y = 5X_1 + 5X_2 + \epsilon.$$

Hence, for this example, the true model is linear and we have  $\beta = (5, 5, 0, \dots, 0)$ . Suppose we use one step of forward step regression to fit the model. Figure 4 shows the meta-confidence set. The confidence intervals for the first two covariates are huge. This correctly reflects the fact that a single step of forward stepwise will sometimes include  $X(1)$  and sometimes will include  $X(2)$ . This variability will not be represented in a method that does not split the data.

Now suppose we do  $k = 5$  steps of forward stepwise. Figure 4 shows the meta-confidence set which now has very narrow intervals. Again, this correctly reflects the fact the model selection is now very stable.

## 6.2 Estimating Variability and Model Stability

Next we consider estimating the amount of variability directly. (A different approach is considered in [Shah and Samworth \(2013\)](#)). For  $1 \leq j \leq d$ , define  $\pi_j = P^n(j \in S)$ , the probability that  $X(j)$  is included in the selected model. If we knew the  $\pi_j$ 's we would have some sense of how variable the process is. In particular, if all the  $\pi_j$ 's are near 0 or 1, then the model selection is stable over splits. Note, again, that  $\pi_j$  contains two sources of variability: the variability of the model selection method and the variability due to splitting.

Using many splits, we can estimate  $\pi_j$ . However, the naive estimator  $\hat{\pi}_j = N^{-1} \sum_{r=1}^N I(j \in S_r)$  may not be consistent. Instead, we use subsampling. We select random subsets of the data  $\mathcal{D}_1, \dots, \mathcal{D}_n$

each of size  $b = n^\beta$  where  $0 < \beta < 1$ . We apply model selection each time to get models  $S_1, \dots, S_N$ . (In principle,  $N = \binom{n}{b}$ . In practice, we simply take a large, random sample of subsamples.) Now the size of these subsets is  $o(n)$  and so will be smaller than  $n/2$ . We expect that this increases the variability and hence will be a conservative estimate of the variability due to model selection and sample splitting. Let

$$\hat{\pi}_j = \frac{1}{N} \sum_{r=1}^N I(j \in S_r).$$

This is a  $U$ -statistic of order  $b$ . It follows from Hoeffding's inequality for  $U$ -statistics and the union bound that

$$P^n(\max_j |\hat{\pi}_j - \pi_j| > \epsilon) \leq 2de^{-2n\epsilon^2/b} = 2de^{-2n^{1-\beta}\epsilon^2}.$$

Hence, as long as  $d = o(e^{n^{1-\beta}})$ , we have that  $\max_j |\hat{\pi}_j - \pi_j| \xrightarrow{P} 0$ .

Subsampling can be combined with sample splitting. These estimates can be based on half the data and be used to help inform the model selection. If a stable model is chosen, then the variability due to sample splitting is less of a concern. If inference is based on the second half of the data, the inferences are valid despite the fact that the model selection procedure is complex and could include predictive measures as well as stability measures.

**Example 19.** *We consider an example with  $n = 50$ ,  $d = 100$ . The covariates are sampled from a standard Gaussian. The true model is linear and is  $Y = 5X(1) + 5X(2) + \epsilon$ . We use  $k$  steps for forward stepwise regression. We estimate the  $\pi_j$ 's by subsampling. Figure 5 shows the estimates of  $\pi_j$  versus  $j$ , for  $k = 1, 2, 5, 10$ . When we underfit, the model is unstable with  $\pi_j \approx .5$  for  $j = 1, 2$ . The model is extremely stable for  $k = 2$ . As  $k$  is allowed to increase, the stability decreases although  $\pi_1$  and  $\pi_2$  remain high as expected.*

### 6.3 Model Averaging (Bagging)

Another way to deal with the variability due to splitting is to average the fitted models over splits. This is usually known as bagging. Here we briefly explain how recent ideas due to [Mentch and Hooker \(2016\)](#) for random forests can be adapted to our case.

As in the previous section, we use subsamples instead of equal splits. Again, let  $\mathcal{D}_1, \dots, \mathcal{D}_N$  be subsamples of size  $n^\beta$  and let  $S_j$  be the model selected by  $\mathcal{D}_j$ . Let  $\hat{\beta}_{S_j}$  be the estimator of  $\beta_{S_j}$ . Define  $\hat{\mu}_j(x) = \hat{\beta}_{S_j}^T x_{S_j}$ . The averaged predictor is

$$\hat{\mu}(x) = \frac{1}{N} \sum_{j=1}^N \hat{\mu}_j(x).$$

[Mentch and Hooker \(2016\)](#) show that, when  $\hat{\mu}$  is a random forest,  $\hat{\mu}(x)$  is an incomplete, infinite order  $U$ -statistic. They derive a central limit theorem for such statistics, so that  $\sqrt{n}(\hat{\mu}(x) - \mathbb{E}(\hat{\mu}(x)))/\xi \rightsquigarrow N(0, 1)$  for some  $\xi$ . Next they consider a LOCO style test. Specifically, they test  $H_0 : \hat{\mu}(x) = \hat{\mu}_{(j)}(x)$

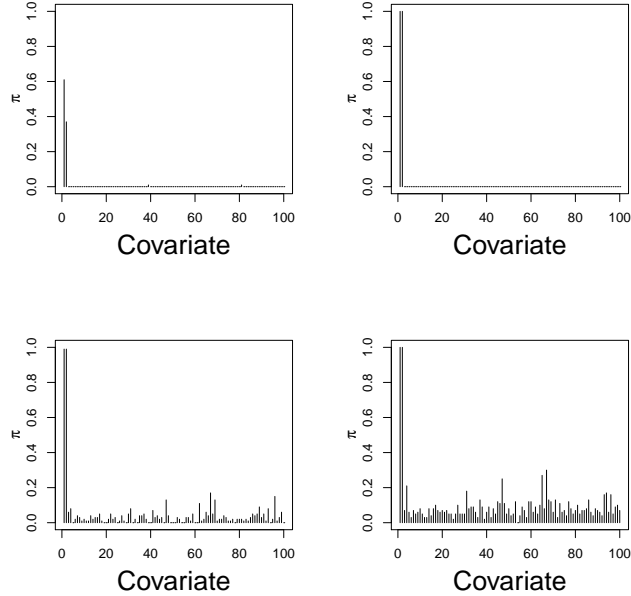


Figure 5: Estimates of  $\pi_j$  versus  $j$  using subsampling.  $n = 50$ ,  $d = 100$ . The models are selected using forward stepwise selection with  $k$  steps. Top left:  $k = 1$ . Top right:  $k = 2$ . Bottom left:  $k = 5$ . Bottom right:  $k = 10$ .

for all  $x$  in some set of test points. Here,  $\hat{\mu}_{(j)}$  is defined as in Section 1.2. Similarly one can define confidence intervals.

We think this approach is very promising. At this time, however, it is not clear if we have the uniform guarantee (1). Also, it is not clear if their CLT holds with increasing  $d$ . For these reasons, we leave further analysis to future work.

## 7 Prediction/Accuracy Tradeoff: Comparing Splitting to Uniform Inference

There is a price to pay for sample splitting: the selected model may be less accurate because only part of the data are used to select the model. Thus, splitting creates gains in accuracy and robustness for inference but with a possible loss of prediction accuracy. We call this the *inference-prediction tradeoff*. In this section we study this phenomenon by comparing splitting with uniform inference (defined below). We use uniform inference for the comparison since this is the any other method we know of that achieves (1). We begin with a simple model where it is feasible to compare splitting with uniform inference.

## 7.1 The Many Means Model (Orthogonal Design)

We start with the *many means problem* which is similar to regression with a balanced, orthogonal design. The data are  $Y_1, \dots, Y_n \sim P$  where  $Y_i \in \mathbb{R}^D$ . Let  $\beta = (\beta(1), \dots, \beta(D))$  where  $\beta(j) = \mathbb{E}[Y_i(j)]$ . In this section, the model  $\mathcal{P}_n$  is the set of probability distributions on  $\mathbb{R}^D$  such that  $\max_j \mathbb{E}|Y(j)|^3 < C$  for some constant  $C$ . We also assume that  $\text{Var}(Y(j))$  is uniformly bounded away from 0 for each  $j$ .

To mimic forward stepwise regression — where we would choose a covariate to maximize correlation with the outcome — we consider choosing  $j$  to maximize the mean. Specifically, we take

$$S \equiv w(Y_1, \dots, Y_n) = \underset{j}{\operatorname{argmax}} \bar{Y}(j) \quad (65)$$

where  $\bar{Y}(j) = (1/n) \sum_{i=1}^n Y_i(j)$ . Our goal is to infer the random parameter  $\beta(S)$ . The number of models is  $D$ . In forward stepwise regression with  $k$  steps and  $d$  covariates, the number of models is  $D = d^k$ . So the reader is invited to think of  $D$  as being very large. We will compare splitting versus non-splitting with respect to three goals: estimation, inference and prediction accuracy.

**Splitting:** In this case we take Let  $\mathcal{D}_1 = \{i : 1 \leq i \leq m\}$  and  $\mathcal{D}_2 = \{i : m+1 \leq i \leq n\}$  where  $m = n/2$ . Then

$$S_m \equiv w(Y_1, \dots, Y_m) = \underset{j}{\operatorname{argmax}} \bar{Y}(j) \quad (66)$$

where  $\bar{Y}(j) = (1/m) \sum_{i=1}^m Y_i(j)$ . The point estimate and confidence interval for the random parameter  $\beta(S)$  are

$$\hat{\beta}_{n/2}(S_m) = \frac{1}{m} \sum_{i=m/2+1}^n Y_i(S)$$

and

$$C_{n/2} = [\hat{\beta}(S) - sz_{\alpha/2}/\sqrt{m}, \hat{\beta}(S) + sz_{\alpha/2}/\sqrt{m}]$$

where  $s^2 = m^{-1} \sum_{i=m/2+1}^n (Y_i(S) - \hat{\beta}(S))^2$ .

**Uniform Inference (Non-Splitting).** By “non-splitting” we mean that the selection rule and estimator are invariant under permutations of the data. In particular, we consider uniform inference which is defined as follows. Let  $\hat{\beta}(s) = n^{-1} \sum_i Y_i(s)$  be the average over all the observations. Let  $S_n = \operatorname{argmax}_s \hat{\beta}(s)$ . Our point estimate is  $\hat{\beta}_n(S)$ . Now define

$$F_n(t) = P^n(\sup_s \sqrt{n} |\hat{\beta}(s) - \beta(s)| \leq t).$$

We can consistently estimate  $F_n$  by the bootstrap:

$$\hat{F}_n(t) = P^n(\sup_s \sqrt{n} |\hat{\beta}^*(s) - \hat{\beta}(s)| \leq t | Y_1, \dots, Y_n).$$

A valid confidence set for  $\beta$  is  $R = \{\beta : \|\beta - \hat{\beta}\|_\infty \leq t/\sqrt{n}\}$  where  $t = \hat{F}_n^{-1}(1 - \alpha)$ . Because this is uniform over all possible models (that is, over all  $s$ ), it also defines a valid confidence interval for a randomly selected coordinate. In particular, we can define

$$C_n = [\hat{\beta}(S) - t/\sqrt{n}, \hat{\beta}(S) + t/\sqrt{n}]$$

Both confidence intervals satisfy (1).

We now compare  $\hat{\beta}_{n/2}(S_m)$  and  $C_{n/2}$  with  $\hat{\beta}_n(S_n)$  and  $C_n$ . The reader should keep in mind that, in general,  $S_m$  might be different than  $S_n$  and hence,  $\beta(S_m) \neq \beta(S_n)$ . The two procedures might be estimating different parameters. We discuss that issue shortly.

**Estimation.** First we consider estimation accuracy.

**Lemma 20.** *For the splitting estimator:*

$$\sup_{P \in \mathcal{P}_n} \mathbb{E} |\hat{\beta}(S) - \beta(S)| \preceq n^{-1/2}. \quad (67)$$

*For non-splitting we have*

$$\inf_{\hat{\beta}} \sup_{P \in \mathcal{P}_n} \mathbb{E} |\hat{\beta}(S) - \beta(S)| \succeq \sqrt{\frac{\log D}{n}}. \quad (68)$$

*We also have for splitting that*

$$\sup_{w \in \mathcal{W}_n} \sup_{P \in \mathcal{P}_n} \mathbb{E} |\hat{\beta}(S) - \beta(S)| \preceq n^{-1/2} \quad (69)$$

*and for non-splitting*

$$\inf_{\hat{\beta}} \sup_{w \in \mathcal{W}_n} \sup_{P \in \mathcal{P}_n} \mathbb{E} |\hat{\beta}(S) - \beta(S)| \succeq \sqrt{\frac{\log D}{n}}. \quad (70)$$

Thus, the splitting estimator converges at a  $n^{-1/2}$  rate. Non-splitting estimators have a slow rate, even with the added assumption of Normality. (Of course, the splitting estimator and non-splitting estimator may in fact be estimating different randomly chosen parameters. We address this issue when we discuss prediction accuracy.)

**Remark:** Similar bounds may be obtained on the size of confidence intervals.

*Proof.* The upper bounds are obvious. The lower bound (68) is from Section 4 in [Sackrowitz and Samuel-Cahn \(1986\)](#). We now show (70). Let  $\hat{\beta} = g(Y)$  be any estimator where  $Y = (Y_1, \dots, Y_n)$ . Given any  $Y$  and any  $w(Y)$ ,  $\hat{\beta}$  provides an estimate of  $\beta(J)$  where  $J = w(Y)$ . Let  $w_j$  be such that  $w_j(X) = j$ . Then define  $\hat{\beta} = (g(Y, w_1(Y)), \dots, g(Y, w_D(Y)))$ . Let  $w_0(Y) = \arg\max_j |\beta(j) - \hat{\beta}(j)|$ . Then  $\mathbb{E}[|\hat{\beta}(J) - \beta(J)|] = \mathbb{E}[||\hat{\beta} - \beta||_\infty]$ . Let  $P_0$  be multivariate Normal with mean  $(0, \dots, 0)$  and identity covariance. For  $j = 1, \dots, D$  let  $P_j$  be multivariate Normal with mean  $\mu_j = (0, \dots, 0, a, 0, 0)$  and identity covariance where  $a = \sqrt{\log D / (16n)}$ . Then

$$\begin{aligned} \inf_{\hat{\beta}} \sup_{w \in \mathcal{W}_n} \sup_{P \in \mathcal{P}_n} \mathbb{E}[|\hat{\beta}(J) - \beta(J)|] &\geq \inf_{\hat{\beta}} \sup_{P \in \mathcal{M}} \mathbb{E}[|\hat{\beta}(J) - \beta(J)|] \\ &= \inf_{\hat{\beta}} \sup_{P \in \mathcal{M}} \mathbb{E}[||\hat{\beta} - \beta||_\infty] \end{aligned}$$

where  $J = w_0(Y)$  and  $M = \{P_0, P_1, \dots, P_D\}$ . It is easy to see that

$$\text{KL}(P_0, P_j) \leq \frac{\log D}{16n}$$

where KL denotes the Kullback-Leibler distance. Also,  $\|\mu_j - \mu_k\|_\infty \geq a/2$  for each pair.

By Theorem 2.5 of [Tsybakov \(2009\)](#),

$$\inf_{\hat{\beta}} \sup_{P \in M} \mathbb{E}[\|\hat{\mu} - \mu\|_\infty] \geq \frac{a}{2}$$

which completes the proof. ■

**Inference.** Now we turn to inference. For splitting, we use the usual Normal interval  $C = [\hat{\beta}(S) - z_\alpha s / \sqrt{m}, \hat{\beta}(S) + z_\alpha s / \sqrt{m}]$  where  $s^2$  is the sample variance from  $\mathcal{D}_2$ . We then have, as a direct application of the one-dimensional Berry-Esseen theorem, that:

**Lemma 21.** *Let  $C$  be the splitting-based confidence set. Then,*

$$\inf_{P \in \mathcal{P}_n} P^n(\beta(S) \in C) = 1 - \alpha - \frac{c}{\sqrt{n}} \quad (71)$$

for some  $c$ . Also,

$$\sup_{P \in \mathcal{P}_n} \mathbb{E}[\nu(C)] \preceq n^{-1/2}. \quad (72)$$

**Lemma 22.** *Let  $C$  be the uniform confidence set. Then,*

$$\inf_{P \in \mathcal{P}_n} P^n(\beta(S) \in C) = 1 - \alpha - \left( \frac{c(\log D)^7}{n} \right)^{1/6} \quad (73)$$

for some  $c$ . Also,

$$\sup_{P \in \mathcal{P}_n} \mathbb{E}[\nu(C)] \succeq \sqrt{\frac{\log D}{n}}. \quad (74)$$

The proof is a straightforward application of results in [Chernozhukov et al. \(2013, 2014\)](#).

We thus see that the splitting method has better coverage and narrower intervals, although we remind the reader that the two methods may be estimating different parameters.

**Can We Estimate the Law of  $\hat{\beta}(S)$ ?** An alternative non-splitting method to uniform inference is to estimate the law  $F_n$  of  $\sqrt{n}(\hat{\beta}(J) - \beta(J))$ . But we show that the law of  $\sqrt{n}(\hat{\beta}(S) - \beta(S))$  cannot

be consistently estimated even if we assume that the data are Normally distributed and even if  $D$  is fixed (not growing with  $n$ ). This was shown for fixed population parameters in [Leeb and Pötscher \(2008\)](#). We adapt their proof to the random parameter case in the following lemma.

**Lemma 23.** *Suppose that  $Y_1, \dots, Y_n \sim N(\beta, I)$ . Let  $\psi_n(\beta) = P^n(\sqrt{n}(\hat{\beta}(S) - \beta(S)) \leq t)$ . There is no uniformly consistent estimator of  $\psi_n(\beta)$ .*

*Proof.* We use a contiguity argument like that in [Leeb and Pötscher \(2008\)](#). Let  $Z_1, \dots, Z_D \sim N(0, 1)$ . Note that  $\hat{\beta}(j) \stackrel{d}{=} \beta(j) + Z_j/\sqrt{n}$ . Then

$$\begin{aligned} \psi_n(\beta) &= P^n(\sqrt{n}(\hat{\beta}(S) - \beta(S)) \leq t) = \sum_j P^n(\sqrt{n}(\hat{\beta}(j) - \beta(j)) \leq t, \hat{\beta}(j) > \max_{s \neq j} \hat{\beta}_s) \\ &= \sum_j P^n(\max_{s \neq j} Z_s + \sqrt{n}(\beta(s) - \beta(j)) < Z_j < t) = \sum_j \Phi(A_j) \end{aligned}$$

where  $\Phi$  is the  $d$ -dimensional standard Gaussian measure and

$$A_j = \left\{ \max_{s \neq j} Z_s + \sqrt{n}(\beta(s) - \beta(j)) < Z_j < t \right\}.$$

Consider the case where  $\beta = (0, \dots, 0)$ . Then

$$\psi_n(0) = D \Phi(\max_{s \neq 1} Z_s < Z_1 < t) \equiv b(0).$$

Next consider  $\beta_n = (a/\sqrt{n}, 0, 0, \dots, 0)$  where  $a > 0$  is any fixed constant. Then

$$\begin{aligned} \psi(\beta_n) &= \Phi((\max_{s \neq 1} Z_s) - a < Z_1 < t) \\ &\quad + \sum_{j=2}^D \Phi(\max\{Z_1 + a, Z_2, \dots, Z_{j-1}, Z_{j+1}, \dots, Z_D\} < Z_j < t) \\ &\equiv b(a). \end{aligned}$$

Suppose that  $\hat{\psi}_n$  is a consistent estimator of  $\psi_n$ . Then, under  $P_0$ ,  $\hat{\psi}_n \xrightarrow{P} b(0)$ . Let  $P_n = N(\beta_n, I)$  and  $P_0 = N(0, I)$ . It is easy to see that  $P_0^n(A_n) \rightarrow 0$  implies that  $P_n^n(A_n) \rightarrow 0$  so that  $P_n$  and  $P_0$  are contiguous. So, by Le Cam's first lemma, under  $P_n$ , we also have that  $\hat{\psi}_n \xrightarrow{P} b(0)$ . But  $b(0) \neq b(a)$ , which contradicts the assumed consistency of  $\hat{\psi}_n$ . ■

**Prediction Accuracy.** Now we discuss prediction accuracy which is where splitting pays a price. The idea is to identify a population quantity  $\theta$  that model selection is implicitly targeting and compare splitting versus non-splitting in terms of how well they estimate  $\theta$ . The purpose of model selection in regression is to choose a model with low prediction error. So, in regression, we might take  $\theta$  to be the prediction risk of the best linear model with  $k$  terms. In our many-means model, a natural analog of this is the parameter  $\theta = \max_j \beta(j)$ .



We have the following lower bound, which applies over all estimators both splitting and non-splitting. For the purposes of this lemma, we use Normality. Of course, the lower bound is even larger if we drop Normality.

**Lemma 24.** *Let  $Y_1, \dots, Y_n \sim P$  where  $P = N(\beta, I)$ ,  $Y_i \in \mathbb{R}^D$ , and  $\beta \in \mathbb{R}^D$ . Let  $\theta = \max_j \beta(j)$ . Then*

$$\inf_{\hat{\theta}} \sup_{\beta} E[(\hat{\theta} - \theta)^2] \geq \frac{2 \log D}{n}.$$

*Proof.* Let  $P_0 = N(\mu_0, \frac{1}{n}I_D)$ , where  $\mu_0 = 0$ , and for  $j = 1, \dots, D$  let  $P_j = N(\mu_j, \frac{1}{n}I_D)$ , where  $\mu_j$  is the  $D$ -dimensional vector with 0 entries except along the  $j$ th coordinate, which takes the value  $\sqrt{c \frac{\log D}{n}}$ , where  $0 < c < 1$ . Consider the mixture  $\bar{P} = \frac{1}{D} \sum_{j=1}^D P_j$ . Then, letting  $\theta_j$  and  $\theta_0$  be the largest coordinates of  $\mu_j$  and  $\mu_0$  respectively, we have that  $\|\theta_j - \theta_0\|^2 = \frac{c \log D}{n}$  for all  $j$ . Next, some algebra yields that the  $\chi^2$  distance between  $P_0$  and the mixture  $\bar{P} = \frac{1}{D} \sum_{j=1}^D P_j$  is  $\frac{1}{D} e^{c \log D} - \frac{1}{D}$ , which vanishes as  $D$  tends to  $\infty$ . Since this is also an upper bound on the squared total variation distance between  $P_0$  and  $\bar{P}$ , the result follows from an application of Le Cam Lemma. ■

To understand the implications of this result, let us write

$$\hat{\beta}(S) - \theta = \underbrace{\hat{\beta}(S) - \beta(S)}_{L_1} + \underbrace{\beta(S) - \theta}_{L_2}. \quad (75)$$

The first term,  $L_1$ , is the focus of most research on post-selection inference. We have seen it is small for splitting and large for non-splitting. The second term takes into account the variability due to model selection which is often ignored. Because  $L_1$  is of order  $n^{-1/2}$  for splitting, and the because the sum is of order  $\sqrt{\log D/n}$  it follows that splitting must, at least in some cases, pay a price by have  $L_2$  large. In regression, this would correspond to the fact that, in some cases, splitting leads to models with lower predictive accuracy.

Of course, these are just lower bounds. To get more insight, we consider a numerical example. Figure (6) shows a plot of the risk of  $\hat{\beta}(S) = \bar{Y}(S)$  for  $n$  (non-splitting) and  $n/2$  (splitting). In this example we see that indeed, the splitting estimator suffers a larger risk. In this example,  $D = 1,000$ ,  $n = 100$ , and  $\beta = (a, 0, \dots, 0)$ . The horizontal axis is  $a$  which is the gap between the largest and second largest mean.

To summarize: splitting gives more precise estimates and coverage for the selected parameter than non-splitting (uniform) inference. But the two approaches can be estimating different parameters. This manifests itself by the fact that splitting can lead to less precise estimates of the population parameter  $\theta$ . In the regression setting, this would correspond to the fact that splitting the data can lead to selecting models with poorer prediction accuracy.

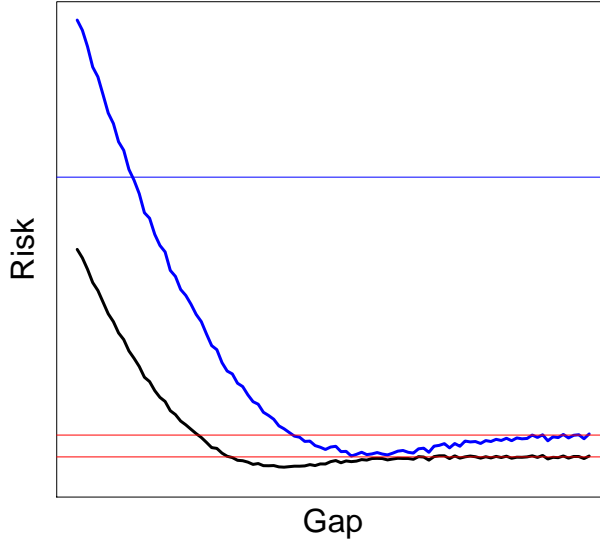


Figure 6: Horizontal axis: the gap  $\beta_{(1)} - \beta_{(2)}$ . Blue risk: risk of splitting estimator. Black line: risk of non-splitting estimator.

## 7.2 Back To General Regression

Now we return to regression without an orthogonal design. We restrict ourselves to selectors with  $|S| = k$  where  $k$  is fixed.

For splitting methods, the question of coverage accuracy has been addressed in Sections 3 and 4. A lower bound on the estimation accuracy  $\sup_{P \in \mathcal{P}'_n} \mathbb{E} \|\hat{\beta}_S - \beta_S\|_\infty$  without splitting has not, to the best of our knowledge, been obtained. This is the focus of current work. We conjecture that the lower bound is  $\sqrt{k \log d/n}$  for non-splitting. For splitting the upper bound is  $\sqrt{\log k/n}$ .

For inference, splitting has accuracy  $n^{-1/2}$ . For non-splitting the question is complicated because no non-splitting method has been shown to satisfy (1) as far as we know. One possibility is to use the bootstrap. We define  $t_\alpha$  by

$$P\left(\sqrt{n} \sup_{S \in \mathcal{S}_k} \|\hat{\beta}_S^* - \hat{\beta}_S\|_\infty > t_\alpha \mid Z_1, \dots, Z_n\right) = \alpha$$

where  $\mathcal{S}_k$  is the set of all models of size  $k$ . The set  $\{\beta_S : \|\hat{\beta}_S - \beta_S\|_\infty \leq t_\alpha/\sqrt{n}\}$  would then have correct asymptotic coverage. It would also be robust to the selection rule. But there are two difficulties here. First, the confidence set would likely be very large. Second, computing  $\sup_{S \in \mathcal{S}_k} \|\hat{\beta}_S^* - \hat{\beta}_S\|_\infty$  is not feasible. If we restrict to Normal distributions with constant, known variance, then the calculation is feasible as shown in [Berk et al. \(2013\)](#). However, this takes us out of our assumption-free framework.

In summary, we do not know of a viable method other than sample splitting in the assumption-free framework.

Now let us bound the increase in prediction risk due to sample splitting. Here, we focus on  $\ell_1$  regression. Following Greenshtein et al. (2004), we define the best,  $\ell_1$ -sparse linear predictor  $\beta_*$  by

$$\mathbb{E}(Y - \beta_*^T X)^2 = \inf_{\beta \in B(L)} \mathbb{E}[(Y - \beta^T X)^2]$$

where

$$B_L = \{\beta : \|\beta\|_1 \leq L\}.$$

Thus,  $\beta_*$  is the population version of the lasso estimator.

The next theorem bounds the excess risk of the lasso. It is a slight refinement of a result in Greenshtein et al. (2004).

**Theorem 25.** *Let  $Z = (Y, X)$ . Assume that  $|Y| \leq B$  and  $\max_j |X(j)| \leq B$ . Let*

$$\beta_* = \underset{\|\beta\|_1 \leq L}{\operatorname{argmin}} r(\beta)$$

where  $r(\beta) = \mathbb{E}(Y - \beta^T X)^2$ . Thus,  $\beta_*$  is the best, sparse linear predictor (in the  $\ell_1$  sense). Let  $\hat{\beta}$  be the lasso estimator:

$$\hat{\beta} = \underset{\|\beta\|_1 \leq L}{\operatorname{argmin}} \hat{r}(\beta)$$

where  $\hat{r}(\beta) = n^{-1} \sum_{i=1}^n (Y_i - X_i^T \beta)^2$ . With probability at least  $1 - \delta$ ,

$$r(\hat{\beta}) \leq r(\beta_*) + \sqrt{\frac{16(L+1)^4 B^2}{n} \log \left( \frac{\sqrt{2} d}{\sqrt{\delta}} \right)}.$$

*Proof.* Let  $Z = (Y, X)$  and  $Z_i = (Y_i, X_i)$ . Define  $\gamma \equiv \gamma(\beta) = (-1, \beta)$ . Then

$$r(\beta) = \mathbb{E}(Y - \beta^T X)^2 = \gamma^T \Lambda \gamma$$

where  $\Lambda = \mathbb{E}[ZZ^T]$ . Note that  $\|\gamma\|_1 = \|\beta\|_1 + 1$ . Let  $\mathcal{B} = \{\beta : \|\beta\|_1 \leq L\}$ . The training error is

$$\hat{r}(\beta) = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^T \beta)^2 = \gamma^T \hat{\Lambda} \gamma$$

where  $\hat{\Lambda} = \frac{1}{n} \sum_{i=1}^n Z_i Z_i^T$ . For any  $\beta \in \mathcal{B}$ ,

$$\begin{aligned} |\hat{r}(\beta) - r(\beta)| &= |\gamma^T (\hat{\Lambda} - \Lambda) \gamma| \\ &\leq \sum_{j,k} |\gamma(j)| |\gamma(k)| |\hat{\Lambda}(j,k) - \Lambda(j,k)| \leq \|\gamma\|_1^2 \delta_n \leq (L+1)^2 \Delta_n \end{aligned}$$

where  $\Delta_n = \max_{j,k} |\hat{\Lambda}(j,k) - \Lambda(j,k)|$ . So,

$$r(\hat{\beta}) \leq \hat{r}(\hat{\beta}) + (L+1)^2 \Delta_n \leq \hat{r}(\beta_*) + (L+1)^2 \Delta_n \leq r(\beta_*) + 2(L+1)^2 \Delta_n.$$

Note that  $|Z(j)Z(k)| \leq B^2 < \infty$ . By Hoeffding's inequality,

$$P^n(\Delta_n(j,k) \geq \epsilon) \leq 2e^{-n\epsilon^2/(2B^2)}$$

and so, by the union bound,

$$P^n(\Delta_n \geq \epsilon) \leq 2d^2 e^{-n\epsilon^2/(2B^2)} = \delta$$

if we choose  $\epsilon = \sqrt{(4B^2/n) \log \left( \frac{\sqrt{2}d}{\sqrt{\delta}} \right)}$ . Hence,

$$r(\hat{\beta}) \leq r(\beta_*) + \sqrt{\frac{16(L+1)^4 B^2}{n} \log \left( \frac{\sqrt{2}d}{\sqrt{\delta}} \right)}.$$

with probability at least  $1 - \delta$ . ■

It follows that, with probability at least  $1 - \delta$ ,

$$\frac{\hat{r}(\hat{\beta}_{n/2})}{\hat{r}(\hat{\beta}_n)} \leq 1 + \frac{1}{r(L)} \sqrt{\frac{16(L+1)^4 B^2}{n} \log \left( \frac{\sqrt{2}d}{\sqrt{\delta}} \right)}$$

where  $r(L) = r(\beta_*)$ . This is a very conservative bound because it assumes that the risk of  $\hat{\beta}_{n/2}$  is far from  $r(L)$  but that the risk of  $\hat{\beta}_n$  actually achieves the best possible risk  $r(L)$ . The second term is an upper bound on the excess prediction loss due to using only have the data to estimate  $\beta$ .

Suppose that  $\mu$  is not linear. In this case, there exists  $c > 0$  such that  $\lim_{L \rightarrow \infty} r(L) \geq c$  and hence

$$\frac{\hat{r}(\hat{\beta}_{n/2})}{\hat{r}(\hat{\beta}_n)} \preceq 1 + \frac{L^2 \sqrt{\log d}}{\sqrt{n}}.$$

For sparse fits, with  $L = O(1)$ , the excess risk is  $1 + O(\frac{\sqrt{\log d}}{\sqrt{n}})$ . However, if we fit large models with little regularization so that  $L \succeq n^{1/4}$ , then the excess risk, at least under this very conservative scenario, could be non-trivial.

If  $\mu$  is linear, then  $r(L) \rightarrow 0$  as  $L \rightarrow \infty$ . The rate at which  $r(L) \rightarrow 0$  will depend on how well  $\mu$  can be approximated by a sparse model. If  $r(L)$  tends to 0 rapidly, and we use a non-sparse model (large  $L$ ) then the excess risk due to splitting could be large. A more thorough investigation of this phenomena is warranted but is beyond the scope of this paper.

## 8 Comments on Non-Splitting Methods

There are several methods for constructing confidence intervals in high-dimensional regression. The approaches in [Javanmard and Montanari \(2014\)](#); [Nickl et al. \(2013\)](#), are based on de-biasing the lasso estimator. They require the linear model to be correct as well as assumptions on the design. They target the true  $\beta$  which is well-defined since the linear model is assumed to be correct. They do not provide the uniform, assumption-free guarantees.

[Lockhart et al. \(2014\)](#); [Lee et al. \(2016\)](#); [Taylor et al. \(2014\)](#) do not require the linear model to be correct nor do they require design conditions. However, their results only hold for parametric models. Their method works by inverting a pivot.

In fact, inverting a pivot is, in principle, a very general approach. We could even use inversion in the nonparametric framework as follows. For any  $P \in \mathcal{P}$  and any  $j$  define  $t(j, P)$  by

$$P^n(\sqrt{n}|\hat{\beta}_S(j) - \beta_S(j)| > t(j, P)) = \alpha.$$

Note that, in principle,  $t(j, P)$  is known. For example, we could find  $t(j, P)$  by simulation. Now let  $A = \{P \in \mathcal{P} : \sqrt{n}|\hat{\beta}_S(j) - \beta_S(j)| < t(j, P)\}$ . Then  $P^n(P \in A) \geq 1 - \alpha$  for all  $P \in \mathcal{P}$ . Write  $\beta_j(S) = f(P, Z_1, \dots, Z_n)$ . Let  $C = \{f(P, Z_1, \dots, Z_n) : P \in A\}$ . It follows that  $P^n(\beta_j(S) \in C) \geq 1 - \alpha$  for all  $P \in \mathcal{P}$ . Furthermore, we could also choose  $t(j, P)$  to satisfy  $P^n(\sqrt{n}|\hat{\beta}_S(j) - \beta_S(j)| > t(j, P) | E_n) = \alpha$  for any event  $E_n$  which would give conditional confidence intervals if desired.

There are two problems with this approach. First, the confidence sets would be huge. Second, it is not computationally feasible to find  $t(j, P)$  for every  $P \in \mathcal{P}$ . The crucial and very clever observation in [Lee et al. \(2016\)](#) is that if we restrict to a parametric model (typically they assume a Normal model with known, constant variance) then, by choosing  $E_n$  carefully, the conditional distribution reduces, by sufficiency, to a simple one parameter family. Thus we only need to find  $t$  for this one parameter family which is feasible. Unfortunately, the method does not provide confidence guarantees of the form (1) which is the goal of this paper.

[Berk et al. \(2013\)](#) is closest to providing the kind of guarantees we have considered here. But as we discussed in the previous section, it does not seem to be extendable to the assumption-free framework.

None of these comments is meant as a criticism of the aforementioned methods. Rather, we just want to clarify that these methods can give narrower intervals but they are not comparable to our results because we require uniformity over  $\mathcal{P}$ . Also, except for the method of [Berk et al. \(2013\)](#), none of the other methods provide any guarantees over unknown selection rules. A summary of the various methods is in Table 1.

## 9 Numerical Examples

In this section we briefly consider a few illustrative examples. In a companion paper, we provide detailed simulations comparing all of the recent methods that have proposed for inference after

Method	Parameter	Assumptions	Accuracy	Computation	Robust
Debiasing	true $\beta$	Very Strong	$1/\sqrt{n}$	Good	No
Conditional	projection	Strong	?	Good	No
Uniform	projection	weak	$\sqrt{k/n}$	NP hard	Yes
Sample Splitting	projection	none	$\sqrt{\log k/n}$	Easy	Yes
Sample Splitting	LOCO	none	$\sqrt{\log k/n}$	Easy	Yes

Table 1: Different inferential methods. ‘accuracy’ refers to the size of the confidence set. ‘robust’ refers to robustness to model assumptions.

model selection. It would take too much space, and go beyond the scope of the current paper, to include these comparisons here.

We focus on linear models, and in particular on inference for the projected parameter ( $\beta_S$ ) and the LOCO parameter ( $\gamma_S$ ) of Section 1.2. The data are drawn from three distributions:

**Setting A** *Linear and sparse with Gaussian noise.* A linear model with  $\beta_i \sim U[0, 1]$  for  $j = 1, \dots, 5$  and  $\beta_j = 0$  otherwise.

**Setting B** *Additive and sparse with  $t$ -distributed noise.* An additive model with a cubic and a quadratic term, as well as three linear terms, and  $t_5$ -distributed additive noise.

**Setting C** *Non-linear, non-sparse,  $t$ -distributed noise.* The variables from Setting B are rotated randomly to yield a dense model.

In Settings A and B,  $n = 100$  (before splitting); in Setting C  $n = 200$ . In all Settings  $p = 50$  and the noise variance is 0.5. The linear model,  $\hat{\beta}_S$  is selected on  $\mathcal{D}_1$  by lasso with  $\lambda$  chosen using 10-fold cross-validation. For  $\gamma_j$ ,  $\hat{\beta}_{S,j}$  is estimated by reapplying the same selection procedure to  $\mathcal{D}_1$  with the  $j^{\text{th}}$  variable removed. Confidence intervals are constructed using the pairs bootstrap procedure of Section 2 with  $\alpha = 0.05$ .

Figure 7 shows typical confidence intervals for the projection parameter,  $\beta$ , and the LOCO parameter,  $\gamma$ , for one realization of each Setting. Notice that confidence intervals are only constructed for  $j \in S$ . The non-linear term is successfully covered in Setting B, even though the linear model is wrong.

Figure 8 shows the coverage probability for Setting B as a function of  $n$ , holding  $p = 50$  fixed. The coverage for the LOCO parameter,  $\gamma_S$  is accurate even at low sample sizes. The coverage for  $\beta_S$  is low (0.8-0.9) for small sample sizes, but converges to the correct coverage as the sample size grows. This suggests that  $\beta_S$  is an easier parameter to estimate and conduct inference on.

## 10 Conclusion

In this paper we have taken a modern look at inference based on sample splitting. We have also investigated the accuracy of Normal and bootstrap approximations and we have suggested new

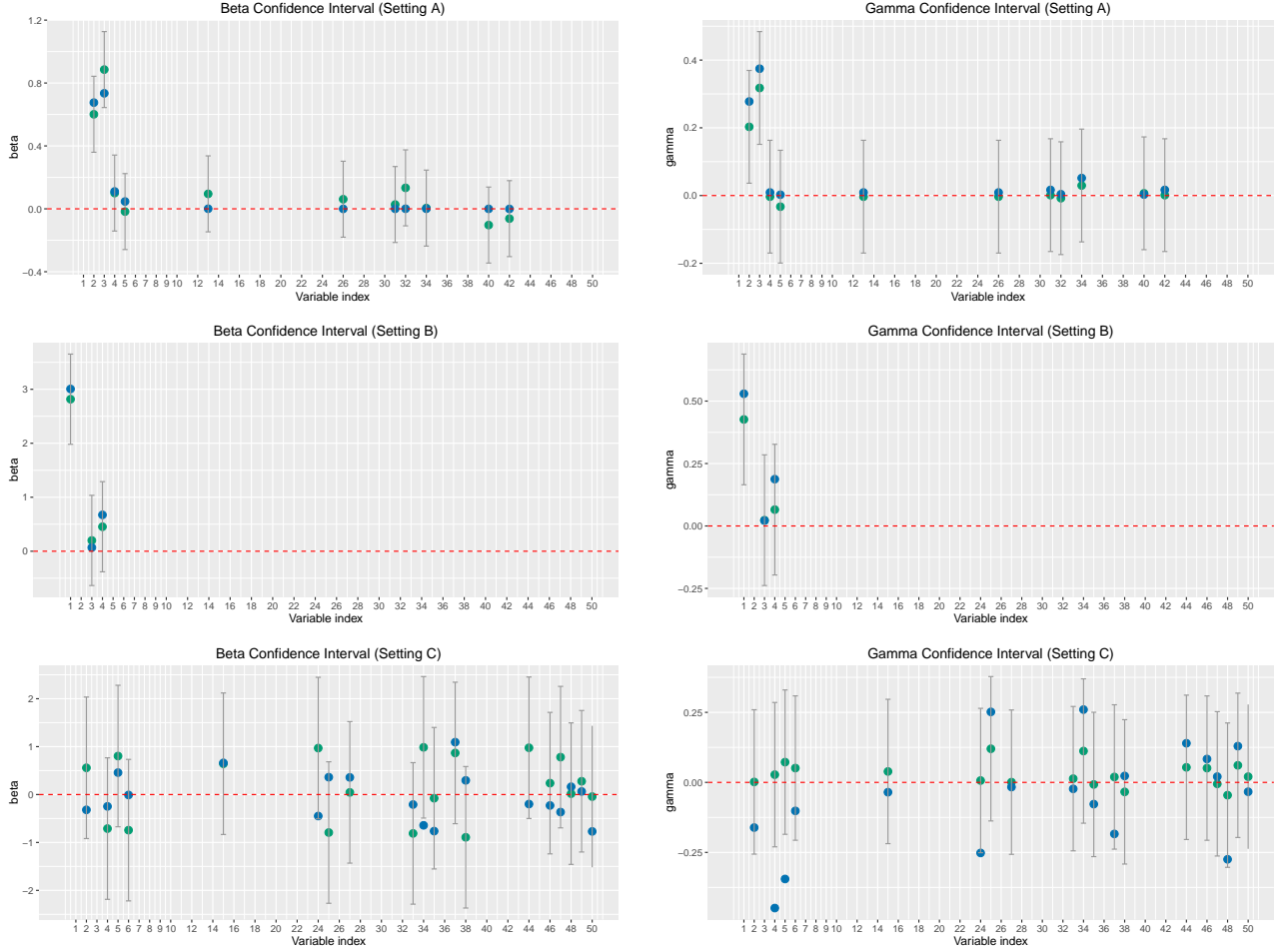


Figure 7: Typical confidence intervals for the projection parameter (left) and the LOCO parameter (right) for Settings A, B, and C. Blue indicates the true parameter value, and green indicates the point estimate from  $\mathcal{D}_2$ . Note that the parameters are successfully covered even when the underlying signal is non-linear ( $X_1$  in Setting B) or dense (Setting C).

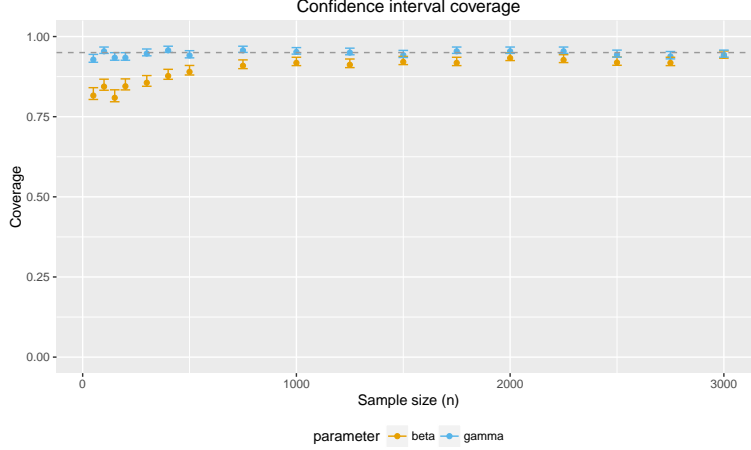


Figure 8: Joint coverage probability of the intervals for  $\beta_S$  and  $\gamma_S$  in Setting B, as sample size  $n$  varies with  $p = 50$  held fixed. The coverage for  $\gamma_S$  is accurate even at low sample sizes, while the coverage for  $\beta_S$  converges more slowly.

parameters for regression.

Despite the fact that sample splitting is an old idea, there remain many open questions. The most pressing is to find a formal mechanism for combining inferences from different splits. Also, the bounds on coverage accuracy — which are of interest beyond sample splitting — are upper bounds. An important open question is to find lower bounds. Also, we are investigating methods for improving the bootstrap rates. For example, the remainder term in the Taylor approximation of  $\sqrt{n}(\hat{\beta}(j) - \beta(j))$  is

$$\frac{1}{2n} \int \int \delta^T H_j((1-u)\psi + u\hat{\psi}) \delta \, du$$

where  $\delta = \sqrt{n}(\hat{\psi} - \psi)$ . By approximating this quadratic term it might be possible to correct the bootstrap distribution. [Pouzo et al. \(2015\)](#) has results for bootstrapping quadratic forms that could be useful here. We saw in Section 5 that a modified bootstrap, that we called the image bootstrap, has very good coverage accuracy even in high dimensions. Future work is needed to compute the resulting confidence set efficiently.

Currently, we are working on a major simulation study to compare the various inferential methods under a variety of conditions. The results of this study will be reported elsewhere. Further work is also needed on methods for combining splits and on the inference-prediction tradeoff.

Finally, we remind the reader that we have taken an assumption-free perspective. If there are reasons to believe in some parametric model then of course sample splitting will be sub-optimal.



## 11 Appendix

We collect some results needed for the proofs. The results are from [Chernozhukov et al. \(2015, 2014\)](#); [Nazarov \(2003\)](#). However, our statement of the results is slightly different than in the original papers. The reason for this is that we need to keep track of some constants in the proofs that affect our rates.

The following anti-concentration result for the maxima of Gaussian vectors follows from Lemma A.1 in [Chernozhukov et al. \(2014\)](#).

**Theorem 26** (Anti-concentration of Gaussian maxima). *Let  $(X_1 \dots, X_p)^\top$  be a centered Gaussian vector in  $\mathbb{R}^p$  with  $\sigma_j^2 = \mathbb{E}[X_j^2] > 0$  for all  $j = 1, \dots, p$ . Moreover, let  $\underline{\sigma} = \min_{1 \leq j \leq p} \sigma_j$ . Then, for any  $y \in \mathbb{R}$  and  $a > 0$*

$$P^n(\max_j |X_j| \leq y + a) - P^n(\max_j |X_j| \leq y) \leq \frac{a}{\underline{\sigma}} \left( \sqrt{2 \log 2k} + 2 \right).$$

The following high-dimensional central limit theorem follows from Proposition 2.1 in [Chernozhukov et al. \(2014\)](#) and Theorem 26. Notice that we have kept the dependence on the minimal variance explicit.

**Theorem 27** (Berry-Esseen bound for simple convex sets). *Let  $X_1, \dots, X_n$  be i.i.d. centered random vectors in  $\mathbb{R}^p$ . Let  $S_n^X = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$  and, similarly, let  $S_n^Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$ , where  $Y_1, \dots, Y_n$  are independent vectors with  $Y_i \sim N_p(0, \mathbb{E}[X_1^\top X_i])$ . Let  $\mathcal{A}$  be the collection of all polyhedra  $A$  in  $\mathbb{R}^p$  of the form*

$$A = \left\{ x \in \mathbb{R}^d : v^\top x \leq t_v, v \in \mathcal{V}(\mathcal{A}) \right\}$$

*where  $\mathcal{V}(\mathcal{A}) \subset \mathbb{R}^p$  is a set of  $m$  points of unit norm, with  $m \leq (np)^d$  for some constant  $d > 0$ , and  $(t_v : v \in \mathcal{V}(\mathcal{A}))$  is a set of  $m$  positive numbers. For each  $i = 1, \dots, n$  let*

$$\tilde{X}_i = (\tilde{X}_{i1}, \dots, \tilde{X}_{im})^\top = \left( v^\top X_i, v \in \mathcal{V}(\mathcal{A}) \right).$$

*Assume that the following conditions are satisfied:*

$$(M1') \quad n^{-1} \sum_{i=1}^n \mathbb{E} \left[ \tilde{X}_{ij}^2 \right] \geq \underline{\sigma}^2, \text{ for all } j = 1, \dots, m;$$

$$(M2') \quad n^{-1} \sum_{i=1}^n \mathbb{E} \left[ B |\tilde{X}_{ij}|^{2+k} \right] \leq B_n^k, \text{ for all } j = 1, \dots, m \text{ and } k = 1, 2;$$

$$(E1') \quad \mathbb{E} \left[ \exp \left( |\tilde{X}_{i,j}| / B_n \right) \right] \leq 2, \text{ for } i = 1, \dots, n \text{ and } k = 1, 2.$$

*Then, there exists a constant  $C > 0$  depending only on  $d$  such that*

$$\sup_{A \in \mathcal{A}} \left| P^n(S_n^X \in A) - P^n(S_n^Y \in A) \right| \leq \frac{C}{\underline{\sigma}} \left( \frac{B_n^2 \log^7(pn)}{n} \right)^{1/6}.$$

Finally, we will use the following comparison theorem for the maxima of Gaussian vectors, obtained by using Theorem 26 instead of Corollary 1 in Chernozhukov et al. (2015) in the proof of Theorem 2 in Chernozhukov et al. (2015). Notice that we have kept the dependence on the minimal variance explicit. Also, we have eliminated the dependence on the maximum variance by noting that the anti-concentration result in Nazarov (2003) does not depend on the maximum variance.

**Theorem 28** (Gaussian comparison). *Let  $X \sim N_p(0, \Sigma_X)$  and  $Y \sim N_p(0, \Sigma_Y)$  with*

$$\Delta = \max_{i,j} |\Sigma_X(j, k) - \Sigma_Y(j, k)|$$

*Let  $\underline{\sigma} = \min_j \Sigma_X(j, j) \vee \Sigma_Y(j, j)$ . Then, there exists a universal constant  $C > 0$  such that*

$$\sup_{t \in \mathbb{R}} \left| P^n(\max_j X_j \leq t) - P^n(\max_j Y_j \leq t) \right| \leq C \frac{\Delta^{1/3} (2 \log p)^{1/3}}{\underline{\sigma}^{1/3}}.$$

## References

- Andreas Anastasiou and Robert E Gaunt. Multivariate normal approximation of the maximum likelihood estimator via the delta method. *arXiv preprint arXiv:1609.03970*, 2016.
- Andreas Anastasiou and Christophe Ley. New simpler bounds to assess the asymptotic normality of the maximum likelihood estimator. *arXiv preprint arXiv:1508.04948*, 2015.
- Andreas Anastasiou and Gesine Reinert. Bounds for the normal approximation of the maximum likelihood estimator. *arXiv preprint arXiv:1411.2391*, 2014.
- G. Barnard. Discussion of “cross-validatory choice and assessment of statistical predictions” by stone. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–135, 1974.
- V Bentkus. Lower bounds for the rate of convergence in the central limit theorem in banach spaces. *Lithuanian Mathematical Journal*, 25(4):312–320, 1985.
- Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.
- Andreas Buja, Richard Berk, Lawrence Brown, Edward George, Emil Pitkin, Mikhail Traskin, Linda Zhao, and Kai Zhang. Models as approximations—a conspiracy of random regressors and model deviations against classical inference in regression. *Statistical Science*, 1460, 2015.
- Louis HY Chen and Qi-Man Shao. Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli*, pages 581–599, 2007.
- Victor Chernozhukov, Denis Chetverikov, Kengo Kato, et al. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.
- Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *arXiv preprint arXiv:1412.3661*, 2014.

- Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162(1-2): 47–70, 2015.
- DR Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.
- Miklos Csorgo. *Quantile processes with statistical applications*, volume 42. SIAM, 1983.
- Noureddine El Karoui and Elizabeth Purdom. Can we trust the bootstrap in high-dimension? *UC Berkeley Statistics Department Technical Report*, 2015.
- Julian J Faraway. Data splitting strategies for reducing the effect of model selection on inference. *Ann Arbor*, 1001:48109, 1995.
- Eitan Greenshtein, Ya’Acov Ritov, et al. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.
- John A Hartigan. Using subsample values as typical values. *Journal of the American Statistical Association*, 64(328):1303–1317, 1969.
- Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.
- Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.
- Hannes Leeb and Benedikt M Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(02):338–376, 2008.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *arXiv preprint arXiv:1604.04173*, 2016.
- Ker-Chau Li. Honest confidence regions for nonparametric regression. *The Annals of Statistics*, pages 1001–1008, 1989.
- Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.
- Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 2012.
- Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41, 2016. URL <http://jmlr.org/papers/v17/14-168.html>.
- Alan Miller. *Subset selection in regression*. CRC Press, 1990.
- PAP Moran. Dividing a sample into two parts a statistical dilemma. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 329–333, 1973.
- Frederick Mosteller and John Wilder Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

- Fedor Nazarov. On the maximal perimeter of a convex set in  $\mathbb{R}^n$  with respect to a gaussian measure. In *Geometric aspects of functional analysis*, pages 169–187. Springer, 2003.
- Richard Nickl, Sara van de Geer, et al. Confidence sets in sparse regression. *The Annals of Statistics*, 41(6):2852–2876, 2013.
- Richard R Picard and Kenneth N Berk. Data splitting. *The American Statistician*, 44(2):140–147, 1990.
- Iosif Pinelis and Raymond Molzon. Berry-esseen bounds for general nonlinear statistics, with applications to pearson’s and non-central student’s and hotelling’s. *Arxiv preprint*, 2009.
- Stephen Portnoy. A central limit theorem applicable to robust regression estimators. *Journal of multivariate analysis*, 22(1):24–50, 1987.
- Demian Pouzo et al. Bootstrap consistency for quadratic forms of sample averages with increasing dimension. *Electronic Journal of Statistics*, 9(2):3046–3097, 2015.
- Harold Sackrowitz and Ester Samuel-Cahn. Evaluating the chosen population: a bayes and minimax approach. *Lecture Notes-Monograph Series*, pages 386–399, 1986.
- Rajen D Shah and Richard J Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.
- Qi-Man Shao, Kan Zhang, and Wen-Xin Zhou. Stein’s method for nonlinear statistics: A brief survey and recent progress. *Journal of Statistical Planning and Inference*, 168:68–89, 2016.
- Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 7, 2014.
- Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.
- Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- Larry Wasserman. Discussion: A significance test for the lasso. *Ann. Statist.*, 42(2):501–508, 04 2014. doi: 10.1214/13-AOS1175E. URL <http://dx.doi.org/10.1214/13-AOS1175E>.